Statistical Aspects of Haplotype-Based Association Studies

by
Bevan Emma Huang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2007

Approved:
Danyu Lin, Advisor
Joseph G. Ibrahim, Reader
Kari North, Reader
Fred Wright, Reader
Donglin Zeng, Reader

ABSTRACT

BEVAN EMMA HUANG: Statistical Aspects of Haplotype-Based Association Studies
(Under the Direction of Danyu Lin)

A decade ago, genomewide association studies were proposed as a tool to unravel the genetic basis of complex diseases. It is only now that they are becoming practical realities due to improved technology and reduced genotyping costs. For such studies, the issues of power and efficiency are crucial due to the quantity of markers genotyped and the moderate effect sizes involved.

Haplotype-based analysis incorporates information from multiple markers, and so is potentially more powerful than single-SNP analysis. Unfortunately, not only is it computationally more intensive, but since haplotypes are not directly observed, there exists a major analytical challenge with haplotype association analysis. Several methods are available to infer individual haplotypes from unphased genotype data, but using the inferred haplotypes in the ensuing association analysis can result in biased estimates and reduced power. We investigate the situations for which the disadvantages of the imputation process may outweigh its convenience. In addition, we describe alternatives to imputation which result in efficient haplotype association analysis.

For case-control studies, we develop methods for use in genomewide studies which account for the correlation between SNPs in multiple test correction. Simulation studies based on the HapMap data showed that the proposed method performs well in realistic situations. We applied it to a case-control dataset of 2,300 SNPs to test for association with rheumatoid arthritis.

For quantitative trait loci, we focus on gains in power which may be made via selective genotyping designs, where only those individuals with extreme phenotypes are genotyped. Because selection depends on the phenotype, the resulting data cannot be properly analyzed by standard statistical methods. We provide appropriate

likelihoods for assessing the effects of genotypes and haplotypes on quantitative traits under such designs. We demonstrate that the likelihood-based methods are highly effective in identifying causal variants, and are substantially more powerful than existing methods. We initially consider two practical designs, then extend the methods to a two-phase sampling design. Additionally, we provide methods to test for haplotype-disease association in the presence of covariates. Simulations demonstrate the effectiveness of these likelihood-based methods.

Dedicated to all those who think the jury's still out on science

ACKNOWLEDGMENTS

I would like to thank my advisor, Danyu Lin, for all the time and support he has given me during my time as a graduate student. Our Monday meetings were a valuable part of my experience. I would also like to give thanks to the members of my committee, Dr. Joe Ibrahim, Dr. Kari North, Dr. Fred Wright, and Dr. Donglin Zeng, for being generous with their comments and advice in our discussions.

To: stalling, hobgood
From: bhuang
You two are sweethearts. Thanks for always having chocolate and smiles to brighten my day.

To: bwheeler, achou
From: behuang
friends don't forget to thank friends for all the wonderful times they've had - locopops, loaded questions, angela's nose, brooke's arms, bubble tea and pho and many more. long live the trifeminate!

To: penguingirl
From: felonius
Subject: you blowhard, tobias
you're getting to be almost as good at wasting my life as casey, but i can't really regret any of it when it's full of AD, wii, chipotle and waking you up early. good times, good times. rock lobster!

To: uncwaterpolo

From: no.seven

the real one, not adrian =P playing polo here has been great. scrimmages against dook, swimming in the summer, quarry trips, and of course, the tournies. i'll just have to try to come back over easter sometime.

To: xpippyx, mathid

From: xsillymex

and yuki, silver, ettan, jade and liffey - but as far as i know, they don't know how to use computers. you're the best sister and brother-in-law (and nieces and nephews) in the world, but rest assured, mats, someday your tricks will come back to haunt you.

To: sromoli, ntakarabe, vicicm

From: gypseee

i'm attempting to spread the spin move all over the world. come visit me soon so i can take care of you as well as you always have me. except for when you made me weird, just like you - unh.

To: zzzzkc

From: eeeeee

your poor luck with cars is surpassed only by your inability to estimate distance. just imagine how much less of my life you'd waste if we could fix those two problems!

To: japasahaj

From: dil.emma

I bet you never thought you'd see this day - all of us are done with school! You'll always have more to teach us, though. There are more things in heaven and earth...

To: mpnerd

From: whatemma

there are many things i could mention here, but the one for which i am most thankful is simply the pleasure of your company. Don't work too hard; those clubs don't run themselves!

To: hamer, rhamer

From: the.goddtr

You made North Carolina feel like home. I can only hope to someday show someone as much generosity and hospitality as you have me.

To: simpsonpippam, kchuangmd

From: ehuangy

I know, I know, back in YOUR day, people finished their degrees in seven years, but they had three babies (the third of which was very time-consuming) and walked to work every day in the snow, barefoot, with a bad ankle... uphill both ways... I guess we're just slackers now. You're always there for me, and I always love you.

-e :)

Date: July 7, 2007

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Bon	Bonferroni
ENCODE	ENcyclopedia Of DNA Elements
FWER	Familywise Error Rate
HWE	Hardy-Weinberg Equilibrium
LD	Linkage Disequilbrium
MAF	Minor Allele Frequency
MC	Monte Carlo
ODS	Outcome-Dependent Sampling
RA	Rheumatoid Arthritis
SNP	Single Nucleotide Polymorphism

Chapter 1

Introduction

1.1 Genetic Association Studies

Genetics research has undergone a dramatic transformation in the past decade as improved technology and reduced cost have made high-throughput genotyping possible. Traditionally, linkage analysis has been used to localize disease genes, and has been successful in the study of various Mendelian traits (Jimenez-Sanchez et al. 2001). This methodology is well developed, but it does have some limitations. It is primarily useful for diseases caused by rare variants with high penetrance, and has poor power to detect common alleles with modest effects (Risch 2000; Hirschhorn and Daly 2005). This makes it difficult if not impossible to study complex traits, which may be greatly influenced by common variants and may have complicated interactions with environmental factors (Reich and Lander 2001; Wang et al. 2005).

Fine mapping, or association analysis, can more closely pinpoint the location of a gene, but requires many more markers. Genomewide association studies were proposed a decade ago as a potentially powerful tool to unravel the genetic basis of complex diseases (Risch and Merikangas 1996). However, it is only now that they are becoming practical realities. Genotyping costs have decreased greatly in recent years, to the point where chips containing 100K SNPs, or even 250K have already been used in various studies (Ozaki et al. 2002; Klein et al. 2005; Thomas et al. 2005; Maraganore

et al. 2005). While genomewide association studies are currently in wide proliferation, the methodology to perform the analysis has not kept pace with the collection of data (Thomas et al. 2005).

One resource in particular which has spurred the creation of new methodology is the International HapMap Project, or HapMap (Gibbs et al. 2003; The International Hapmap Consortium 2005), which has provided researchers with a dense SNP map consisting of over 3.5 million validated SNPs. Containing information on location, allele frequency, and linkage disequilibrium, this compilation of data has potential which is yet to be fully realized. Already approaches have been suggested to use the data to explore the distribution of linkage disequilibrium, construct simulation datasets to accurately represent variation in the genome and select SNPs for genotyping in association studies (The International HapMap Consortium 2005).

The promise for future research seems boundless, yet there are still limitations inherent in the data. The HapMap project and current SNP platforms focus on cataloging common SNPs, so single-SNP analysis is not capable of detecting rare causative SNPs. An alternative is haplotype-based analysis, which may be able to do so if the rare SNP is captured by a haplotype (De Bakker et al. 2005). An important question in testing association between SNPs and disease is whether to examine individual SNPs, or consider the haplotypes of multiple markers. The latter is potentially more powerful due to the incorporation of information from multiple markers (Collins et al. 1997; Akey et al. 2001; Morris and Kaplan 2002), but it is also computationally more intensive.

Further, a major analytical challenge with haplotype association analysis is that haplotypes are not directly measured. While this is theoretically possible, it remains too expensive for use in large-scale association studies. Several methods are available to infer individual haplotypes from unphased genotype data (Schaid et al. 2002; Zaykin et al. 2002; Excoffier and Slatkin 1995; Stephens et al. 2001; Niu et al. 2002), but using the inferred haplotypes in the ensuing association analysis can result in biased estimates and reduced power (Kraft et al. 2005). This dissertation considers

aspects of haplotype-disease association mapping. Chapter 2 lays out general notation which will be utilized throughout this dissertation. The remaining chapters propose efficient methods of analysis under a variety of study designs.

1.2 Association Mapping for Dichotomous Traits

Much interest centers on the etiology of complex diseases such as cancer, which are affected by a multitude of genetic and environmental factors. For such traits the case-control design, which is common in classical epidemiologic studies, is a popular approach to studying the role of genes in influencing disease risk. In this case a typical single-SNP analysis consists of comparing the frequency of the three possible genotypes between cases and controls with a standard χ^2 test. This is equivalent to logistic regression upon the genotype. Once we consider latent covariates such as multi-SNP haplotypes, however, more complex methodology is required.

In Chapter 3, we begin by illustrating problems with the simplistic approach of using standard logistic regression on inferred haplotypes. We compare this with maximum likelihood methods, which properly account for phase uncertainty. This approach involves maximization of the observed-data likelihood with respect to all relevant parameters (including haplotype frequencies and disease risks) simultaneously. Through extensive simulation studies we investigate the types of scenarios for which the disadvantages of the imputation process may outweigh its convenience.

In Chapter 4 we consider genomewide scans for haplotype association. We develop a statistically powerful and numerically efficient method for detecting haplotype-disease association in genomewide studies by sliding windows of SNPs over the genome. This consists of an algorithm to calculate a proper likelihood-ratio statistic for any given window of SNPs, along with an accurate Monte Carlo procedure to adjust for multiple testing. Simulation studies based on the HapMap data showed that the proposed method performs well in realistic situations. We applied it to a real case-control dataset of 2,300 SNPs to test for association with rheumatoid arthritis. Several loci

were identified as having possible effects on the disease, none of which would have been detected with existing methods.

1.3 Association Mapping for Quantitative Traits

While case-control studies are commonly used for genomewide association studies, they are not the most efficient design when the trait of interest is continuous. Efficiency and power are important concerns, since disease genes are unlikely to have very large effects on quantitative traits. The need to adjust for multiple testing only adds to the problem, and despite the continuing improvements in genotyping efficiency, it is still highly expensive to genotype a large number of individuals in genomewide association studies. A cost-effective strategy is to preferentially genotype individuals whose trait values deviate from the population mean. Known as selective genotyping, this approach can result in a substantial increase in power (relative to random sampling with the same number of individuals) because much of the genetic information resides in individuals with extreme phenotypes (Laitinen et al. 1997; Slatkin 1999; van Gestel et al. 2000; Xiong et al. 2002; Chen et al. 2005; Cornish et al. 2005; Wallace et al. 2006).

In Chapter 5, we focus on continuous traits and the gains in power which may be made via selective genotyping designs. Because selection depends on the phenotype, the resulting data cannot be properly analyzed by standard statistical methods. We provide appropriate likelihoods for assessing the effects of genotypes and haplotypes on quantitative traits under such designs. We demonstrate that the likelihood-based methods are highly effective in identifying causal variants, and are substantially more powerful than existing methods.

In Chapter 6, we extend the results on selective genotyping to more general designs. Complex traits may be greatly influenced by environment in addition to genetic variants, so it is essential to consider designs which allow for environmental covariates. We consider a two-phase sampling design under a range of conditions, and provide algo-

rithms to maximize the corresponding likelihoods. Simulations show the effectiveness of these likelihood-based methods in comparison to existing approaches.

Chapter 2

Notation

We adopt the notation of Lin et al. (2005) in this dissertation, for both haplotype-disease association analysis, and more complicated models which include other environmental variables. Suppose we have data on n individuals, each genotyped at M biallelic SNPs. At each locus, the two possible alleles are denoted by 0 and 1. Then each haplotype h is a binary word of length M. The total number of possible haplotypes is $L = 2^M$, although the actual number of haplotypes consistent with the observed data is usually much smaller. For $l = 1, \ldots, L$, let h_l denote the lth possible haplotype.

For each individual, the multi-SNP genotype is an ordered sequence of M elements from the set $\{0, 1, 2\}$. Let H denote the individual's diplotype, the pair of haplotypes on the two homologous chromosomes, and let G be the corresponding (unphased) genotype. Note that G is the sum of the two haplotypes, and as such codes the number of '1' alleles at each locus. We write $H = (h_k, h_l)$ if the individual's diplotype consists of haplotypes h_k and h_l. We cannot exactly determine H on the basis of G if the individual is heterozygous at more than one SNP or if any locus genotype is missing.

Let Y be the phenotype of interest (either discrete or continuous), and let X be a set of environmental variables or covariates. We are interested in estimating the effects of H and possibly X on Y. This relationship can be characterized by the

conditional density function $P(Y|X, H; \theta)$ indexed by a set of parameters θ. There are various choices for the association or disease model. Suppose that h^* is the target haplotype of interest. Then in the absence of covariates, we may employ a linear predictor $\alpha + \beta I(h_k = h_l = h^*)$ under a recessive model, $\alpha + \beta[I(h_k = h^*) + I(h_l = h^*) - I(h_k = h_l = h^*)]$ under a dominant model, or $\alpha + \beta[I(h_k = h^*) + I(h_l = h^*)]$ under an additive model, where h_k and h_l are the pair of haplotypes in H, and $I(\mathcal{A})$ takes the value 1 or 0, dependent on whether the event $\mathcal{A}$ is true or false. We consider all three models in our analyses, but focus on the additive model, since it is thought that contributions to disease risk will often be roughly additive (Balding 2006).

Chapter 3

The Use of Inferred Haplotypes in Downstream Analyses

Marchini et al. (2006) provide a comprehensive description of phasing algorithms for inferring individual haplotypes from unphased genotype data. The authors state that an unresolved question is "whether and, if so, how best to use inferred haplotypes in downstream analyses". The question is important since knowing individual haplotypes is rarely an end in itself. Rather, the aim is to approach the gold standard of molecular haplotyping in accuracy, so that the phased haplotypes can be used in further analysis. By treating inferred haplotypes as known quantities, standard statistical methods and computer programs can easily be used for analysis. The convenience of this approach makes it a tempting tactic for haplotype analysis; however, as this chapter shows, there are downsides to simplicity.

3.1 Haplotype Phase Imputation

Phase ambiguity is a kind of missing data, and using inferred haplotypes in downstream analyses is a form of imputation. The voluminous statistical literature on missing data casts light on the potential pitfalls of imputation. In the words of Dempster and Rubin (1983):

The idea of imputation is both seductive and dangerous. It is seductive

> because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial bias.

Marchini et al. (2006) consider several phasing algorithms in their analysis. They point out that all the phasing algorithms assume Hardy-Weinberg Equilibrium (HWE). Even when the general population is in HWE, the case sample and the pooled case-control sample may not be. Thus, the phasing algorithms may produce biased estimation of haplotype distributions with case-control data. The influence of departures from HWE on estimation accuracy depends on the directionality of the disequilibrium. Another possible source of bias is the fact that the phasing algorithms do not acknowledge the selective-sampling feature of the case-control design. Also, the phasing algorithms do not take account of phenotype, which is potentially informative about phase.

The common practice of assigning the most likely diplotype (i.e., the pair of haplotypes with the highest posterior probability) to each individual is intrinsically biased because the most likely diplotype is not necessarily the true diplotype. Consider the simple situation of two SNPs, with the minor and major alleles coded as 1 and 0, respectively, at each SNP site. The genotype is defined as the numbers of minor alleles at the two SNP sites. Haplotype ambiguity arises if and only if an individual is doubly heterozygous, i.e., has the 11 genotype. Both the (10, 01) and (00, 11) diplotypes produce the 11 genotype. There is obviously a problem if all doubly-heterozygous individuals are assigned with the more likely (i.e., the more common) of the two diplotypes, especially when the frequency of the less common diplotype is similar to (although lower than) that of the more common one.

When there exist causal haplotypes, phasing algorithms may incorrectly assign causal haplotypes to individuals without causal haplotypes or reconstruct causal haplotypes as non-causal haplotypes. Consequently, treating inferred haplotypes as true haplotypes in downstream association analyses tends to attenuate the estimated haplotype effects and reduce the power for detecting causal variants. Incorrect haplotype

Table 3.1: Effects of incorrectly assigned haplotypes on risk estimates.

	Diplotype			
	A	B	C	D
(a) True haplotypes				
Cases	500	100	200	200
Controls	250	150	300	300
Odds ratio	3.0	1.0	1.0	—
(b) Inferred haplotypes				
Cases	300	200	200	300
Controls	150	200	300	350
Odds ratio	2.3	1.2	0.8	—

assignments may also induce spurious association for non-causal haplotypes and thus increase false positive results.

For illustration, we consider the diplotype distribution from a hypothetical case-control study shown in Table 3.1(a). With diplotype D as the reference, the estimated odds ratios for diplotypes A, B and C are 3, 1 and 1, respectively. Assume that, for both cases and controls, 20% of the individuals truly with diplotype A are incorrectly assigned with diplotype B, and another 20% are incorrectly assigned with diplotype D, yielding the misclassified distribution shown in Table 3.1(b). Then the estimated odds ratio for diplotype A is reduced from 3 to 2.3, and the estimated odds ratios for diplotypes B and C are changed from 1 to 1.2 and 0.8, respectively. This example demonstrates that treating inferred haplotypes as true haplotypes may bias the estimated effects of causal haplotypes downward and may also bias the estimated effects of non-causal haplotypes away from the null value in either direction. The distortions can be more profound if the misclassification rates differ between cases and controls.

Several simulation studies (Morris et al. 2002; Kraft et al. 2005; Cordell 2006; French et al. 2006) showed that imputation can yield substantial bias of estimated genetic effects, poor coverage of confidence intervals and significant inflation of type I error, especially when the effects are large and phase uncertainty is high. Kraft et al. (2005) compare the performance of several analytic strategies with matched case-control data. These include the most likely haplotype assignment, expectation substitution, and an improper version of multiple imputation. Cordell (2006) extends this list by weighted regression and consider estimation using either the pooled case/control sample, or separately by disease status. They find multiple imputation to be the easiest to implement and extend to more complex models.

French et al. (2006) focuses primarily on weighted logistic regression analysis. They reported bias of the estimated log odds ratios in the range of -0.49 to 0.22, actual type I error of 18% at the 5% nominal significance level and coverage of less than 40% for 95% confidence intervals. Indeed, when the estimator is biased, the coverage of the association confidence intervals will decrease toward 0% and the type I error will increase toward 100% as the sample size increases.

Of the numerous phasing algorithms investigated in these papers, Marchini et al. (2006) conclude that PHASE (v2.1) (Stephens et al. 2001; Stephens and Donnelly 2003; Stephens and Scheet 2005) is among the most accurate. It was also the method chosen to produce haplotypes for the HapMap data, so we focus on its performance as a standard for haplotype imputation. It is a Bayesian approach to haplotype inference that uses coalescent based models to improve accuracy of haplotype estimates for unrelated individuals. The algorithm attempts to cluster groups of similar haplotypes, and recent versions incorporate recombination, so that the clustering may change as one moves along the chromosome.

3.2 Retrospective Likelihood Maximization

In recent years, researchers (Schaid et al. 2002; Epstein and Satten 2003; Spinka et al. 2005; Lin et al. 2005; Lin and Zeng 2006) have developed maximum likelihood methods to properly account for phase uncertainty in association analyses. This approach involves maximizing the observed-data likelihood with respect to all relevant parameters (including haplotype frequencies and disease risks) simultaneously. We compare the results from this method to those which first impute haplotypes with PHASE and then treat the inferred haplotypes as known covariates in prospective logistic regression.

We first estimate the frequencies for all possible haplotypes for cases and controls separately by using the EM algorithm of Excoffier and Slatkin (1995). The number of haplotypes is denoted by K. For $k = 1, \ldots, K$, let h_k denote the kth haplotype and let π_k denote the frequency of h_k in the whole population. The observed data consist of (Y_i, G_i), $i = 1, \ldots, n$, where Y_i and G_i denote the disease status and genotype for the ith subject. We fit a logistic regression model which takes the form:

$$\Pr(Y = 1|H = (h_k, h_l)) = \frac{e^{\alpha+\beta^T Z(h_k,h_l)}}{1 + e^{\alpha+\beta^T Z(h_k,h_l)}},$$

where α pertains to the intercept, β represents log-odd ratios, and $Z(h_k, h_l)$ represents the design matrix encoding additive haplotype effects. This model has been previously described by Lin et al. (2005) to compare a single haplotype h^* to all other haplotypes, in which case $Z(h_k, h_l)$ counts the number of copies of h^* in a given diplotype. We may also incorporate multiple haplotypes into our model by using

$$Z(h_k, h_l) = \begin{bmatrix} I(h_k = h_1) + I(h_l = h_1) \\ \vdots \\ I(h_k = h_r) + I(h_l = h_r) \end{bmatrix},$$

where r is the number of haplotype effects in the model, and $I(\cdot)$ is the indicator function. We use the most frequent haplotype as the reference group in the model unless otherwise specified.

The likelihood should take into account the phase uncertainty in the genotype data as well as the biased sampling of the case-control design. Under the assumption of rare disease and Hardy-Weinberg equilibrium, the likelihood $\prod_{i=1}^{n} \Pr(G_i|Y_i)$ can be shown to be

$$\prod_{i=1}^{n} \frac{\Sigma_{k,l} I[(h_k, h_l) \in S(G_i)] e^{Y_i \beta^T Z(h_k,h_l)} \pi_k \pi_l}{\Sigma_{k,l} e^{Y_i \beta^T Z(h_k,h_l)} \pi_k \pi_l},$$

where $S(G)$ denotes the set of haplotypes compatible with genotype G, and the summation of (k, l) is taken from 1 to K.

To incorporate the constraints that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$, $k = 1, \ldots, K$, into the calculations, we reparametrize the model by defining $\pi_k^* = \pi_k / \pi_K$ and $\nu_k = \log \pi_k^*$, $k = 1, \ldots, K$. Write $\nu = (\nu_1, \ldots, \nu_{K-1})$ and $\theta = (\beta, \nu)$. Then the log-likelihood can be written as

$$l(\theta) = \sum_{i=1}^{n} \log \left[\frac{\sum_{k,l} I\left[(h_k, h_l) \in S(G_i)\right] e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} e^{\theta^T W(Y_i, h_k, h_l)}} \right],$$

where

$$W(Y_i, h_k, h_l) = \begin{bmatrix} Y_i Z(h_k, h_l) \\ I(h_k = h_1) + I(h_l = h_1) \\ \vdots \\ I(h_k = h_{K-1}) + I(h_l = h_{K-1}) \end{bmatrix}.$$

The corresponding score function is

$$U(\theta) = \sum_{i=1}^{n} \left[\frac{\sum_{k,l} I\left[(h_k, h_l) \in S(G_i)\right] W(Y_i, h_k, h_l) e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} I\left[(h_k, h_l) \in S(G_i)\right] e^{\theta^T W(Y_i, h_k, h_l)}} - \frac{\sum_{k,l} W(Y_i, h_k, h_l) e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} e^{\theta^T W(Y_i, h_k, h_l)}} \right],$$

and to obtain the maximum likelihood estimate $\widehat{\theta}$, we solve the score equation $U(\theta) = 0$ by using the Newton-Raphson method. We set the initial value of θ to $\widetilde{\theta} = (0, \widetilde{\nu})$, where $\widetilde{\nu}$ is the maximum likelihood estimate of ν in the pooled sample obtained by the EM algorithm. We test for haplotype-disease association using Wald statistics for individual haplotype effects, which have approximately the χ^2 distribution with 1 degree of freedom.

3.3 Simulation Studies

The maximum likelihood estimators for haplotype effects are unbiased and statistically efficient, which implies that maximum likelihood is the most powerful among all valid methods (Lin and Zeng 2006). The question we seek to address is how much more powerful it is relative to imputation. We performed several simulation studies to assess bias in estimates and standard error, coverage probability, and power when phase is imputed in case-control studies.

We start with the simple case of a two-locus model where there is only one ambiguous genotype, and examine effects of changing linkage disequilibrium between SNPs. The two SNPs were generated with frequencies of 0.3 and 0.4, with values of LD ranging from $D' = 0$ to 1. Haplotype 11 was assumed to be causative under both additive and dominant modes of inheritance, with odds ratios of 1.0, 1.3, 1.5, 1.8 and 2.0. Based on 10,000 simulated datasets of 1,000 cases and 1,000 controls, with no missing data, we confirmed several of the observations from previous studies. Notably, the imputation method had increasing bias, and decreasing coverage probability for the 95% confidence interval as the odds ratio increased. These trends were more pronounced for lower values of LD, where there was more uncertainty in imputation. The maximum-likelihood method was virtually unbiased and had appropriate coverage under all simulation scenarios. In spite of the issues with bias and coverage, the power for the imputation method was very similar to the maximum-likelihood method as long as D' was larger than 0.2.

Our second study mimicked the two-locus model Mul3 of Cordell (2006). We assumed that haplotypes 01 and 10 had odds ratios of 1.2 and 1.4 in reference to haplotypes 00 and 11 with additive mode of inheritance, and we tested whether locus 2 had an effect while allowing an effect at locus 1. Based on 10,000 simulated data sets of 1,000 cases and 1,000 controls with 10% randomly missing genotypes, we obtained power of 65%, 40% and 17% at nominal significance levels of 5%, 1% and 0.1% for the maximum-likelihood method, as compared to 41%, 20% and 6% for the imputation

Table 3.2: Common haplotypes and population frequencies for AGTR1 (French et al. 2006)

Designation	Haplotype	Frequency
A (Reference)	000100000000	0.223
B	110000000000	0.029
C	110000111110	0.051
D	001010000000	0.027
E	001010000001	0.090
F	000100001001	0.029
G	000100000001	0.188
H	010011000000	0.038
I	010011000001	0.032

method.

Most studies contain more than two SNPs, so in our third study, we considered the type I angiotensin receptor (AGTR1) gene of French et al. (2006), for which 12 SNPs were genotyped, with 9 "common" haplotypes listed in Table 3.2. The average pairwise D' (standardized linkage disequilibrium coefficient) is 0.9. We generated case-control data under the third model in their Table III, but we used the more moderate odds ratios of 2.5, 2, 1.5 and 2 for haplotypes D, F, G, and H, respectively. We assigned disease status under the additive mode of inheritance such that the disease prevalence was approximately 2%, and we selected 800 subjects with 3 controls per case. The power of the maximum-likelihood method to detect the effects of haplotypes D, F, G and H was estimated based on 10,000 simulated data sets with 2% randomly missing SNPs. In this study, approximately 75% of individuals had unambiguous diplotypes, and approximately 82% had highest posterior probabilities greater than 0.75. As can be seen in Table 3.3, the maximum-likelihood methods had substantially higher power than imputation for individual haplotypes. This difference is only amplified by the

Table 3.3: Type I error/Power of maximum likelihood (ML) compared to imputation with PHASE. The nominal significance level is 1%.

Haplotype	Odds Ratio	ML	PHASE
D	2.5	62%	50%
E	1.0	1%	2%
F	2.0	49%	39%
G	1.5	42%	24%
H	2.0	50%	32%

fact that using inferred haplotypes resulted in inflated type I error. Thus the power estimates are higher than they would be when scaled down to correct type I error.

The phasing algorithms reviewed by Marchini et al. (2006) are often used to phase larger regions, so it is of interest to assess the performance of the imputation method when testing for haplotype-disease association on a small set of SNPs that is phased within a larger genomic context. To this end, we generated 100 SNPs according to the allelic frequencies and pairwise linkage disequilibrium coefficients of the first 100 SNPs on chromosome 18 of the CEU sample in the HapMap genome-wide data. We performed haplotype analysis on SNPs 60-64. The most common haplotypes of the 5 SNPs are 00000, 00001, 00010, 00100, 00101, 01101, 10000, 10001, 10010, 10100, and 10101 with frequencies of 4.6%, 8.8%, 11.0%, 7.4%, 7.2%, 7.0%, 6.6%, 6.8%, 8.6%, 7.4% and 8.4%, respectively. We assumed that the disease risk was influenced by haplotype 00000 only, with an odds ratio of 3 under the additive mode of inheritance.

The overall disease prevalence was set to approximately 5%, and we selected 300 cases and 300 controls. We assessed the haplotype-disease association on those 5 SNPs, which were phased together with the other 95 SNPs by the PHASE algorithm. It was not computationally feasible to phase 600 subjects all together on 100 SNPs. Thus, we randomly divided the 600 subjects into 6 groups of 50 cases and 50 controls. (We found that phasing cases and controls together provided much better control of type I

Table 3.4: Standardized LD Coefficients (D') for Two Sets of SNPs on Chromosome 18 of the HAPMAP CEU Sample

(a) SNPs 60-64

	61	62	63	64
60	1.0	.86	.28	.68
61		.86	1.0	.84
62			.55	.73
63				.51

(b) SNPs 95-99

	96	97	98	99
95	1.0	1.0	1.0	.96
96		.83	.95	.94
97			.95	.77
98				.94

error than phasing cases and controls separately.) Recently a new modification to the algorithm, fastPHASE (Scheet and Stephens 2006), was released, for which phasing larger groups of SNPs and subjects may be possible. However, this program has lower accuracy than the version we used, so the results presented here would also apply.

We simulated 1000 datasets with 2% randomly missing SNP values. We found that at the nominal significance level of 1%, the imputation method had 63% power to detect the causal haplotype 00000 and type I error of 5%, 3%, 4% and 7% for null haplotypes 00001, 00010, 00100, and 10000 respectively. The maximum-likelihood method, in contrast, had 72% power to detect the causal haplotype, and type I error close to the nominal level. The maximum likelihood estimates had little bias, whereas the imputation method produced bias of –0.33, 0.27, 0.21, 0.26, and 0.30 for the log odds ratios of haplotypes 00000, 00001, 00010, 00100 and 10000, respectively.

In the above study, the LD among the 5 SNPs was not particularly strong; see Table 3.4(a). In a related study, we considered SNPs 95-99, which had very high LD; see Table 3.4(b). The most common haplotypes of SNPs 95-99 are 00000, 00001, 01000, 01001, 01100, 01111, 10000, and 10001, with frequencies of 39.7%, 20.8%, 2%, 1.3%, 1.8%, 13.8%, 12.9%, and 5.4%, respectively. We assumed that 10001 is the causal haplotype with an odds ratio of 2.5 under the additive mode of inheritance. The rest of the simulation set-up was the same as in the previous study. The imputation

method had 83% power to detect the causal haplotype and type I error of 2% and 4% for null haplotypes 00001 and 10000 at the nominal significance level of 1% and produced bias of –0.15, 0.12, and 0.14 for the log odds ratios of haplotypes 10001, 00001, and 10000. On the other hand, the maximum-likelihood method had 92% power to detect the causal haplotype and provided accurate control of type I error and unbiased estimates of haplotype effects.

3.4 Discussion

Our studies obviously do not encompass all possible scenarios. Thus, the results do not imply that imputation is always bad, but rather that it can be considerably less powerful than maximum likelihood while providing biased estimates of genetic effects and poor control of type I error in practical situations. The problems tend to be more severe when there is greater uncertainty in reconstructed haplotypes.

Our simulation studies were focused on single imputation, which is the most common practice. Some alternative procedures have been proposed, including multiple imputation, expectation substitution, and weighted logistic regression (Kraft et al. 2005; Cordell 2006; French et al. 2006). Those procedures are not theoretically valid either (for many of the reasons mentioned previously) and may perform poorly. In particular, the versions of multiple imputation that have been proposed are improper because they fail to account for phenotype and case-control sampling. Proper multiple imputation would provide a good approximation to maximum likelihood.

Mensah et al. (2007) consider corrections to haplotype imputation to account for the uncertainty in inference. They compare the performance of three methods: (1) treating the PHASE-inferred haplotypes as known quantities; (2) weighting each haplotype pair by its posterior probability; and (3) considering each sampled reconstruction as being an imputation of the true unknown haplotypes, and constructing an estimate based on multiple reconstructions. They show that the latter two can reduce bias and improve coverage probabilities over the first approach when haplotypes

are inferred separately for cases and controls. When haplotypes are inferred for the combined sample, though, the three approaches are equivalent. These improvements to the haplotype imputation process are promising, yet Mensah et al. (2007) admit that they consider only a very limited set of simulations, and provide no discussion of power. Indeed, no method can be more powerful than maximum likelihood while providing the same control of type I error, although some methods may approximate maximum likelihood well under certain circumstances.

Chapter 4

Detecting Haplotype Effects in Genomewide Association Studies

In the previous chapter, we considered haplotype association analysis for a small set of markers. However, as genomewide association studies are becoming widespread in practice, the more relevant question is that of analysis in large-scale association studies. Genotyping costs have decreased greatly in recent years, to the point where chips containing 100K SNPs, or even 250K have already been used in various studies (Ozaki et al. 2002; Klein et al. 2005; Thomas et al. 2005; Maragenore et al. 2005), and investigations of larger numbers of SNPs loom in the near future. However, the methodology to perform the analysis has not kept pace with the collection of data (Thomas et al. 2005).

A major analytical challenge is that haplotypes are not directly measured. Several methods are available to infer individual haplotypes from unphased genotype data (e.g., Excoffier and Slatkin 1995; Stephens et al. 2001; Niu et al. 2002). As seen in the previous chapter, using the inferred individual haplotypes in the ensuing association analysis can result in biased estimates and reduced power. A few methods have been proposed to properly account for phase uncertainty in the association analysis (Zhao et al. 2003; Stram et al. 2003; Epstein and Satten 2003; Lin et al. 2005), but these are all focused on the analysis of a single candidate gene.

In this chapter, we provide a computationally efficient and statistically powerful method for detecting haplotype-disease association in genomewide studies. We consider sliding windows of adjacent SNPs; see Mathias et al. (2006) and the references therein. Within each window, we use an efficient and stable algorithm to calculate a likelihood-ratio test statistic that properly accounts for phase uncertainty and case-control sampling. The windows may be overlapping or non-overlapping, and the window sizes may be fixed or variable. We allow exhaustive testing, which considers all possible windows up to a certain size and thus encompasses single-SNP analysis.

The number of tests can be very large, particularly in the case of exhaustive testing. It is common to use the Bonferroni correction to adjust for multiple testing, but this is overly conservative, especially for overlapping windows and exhaustive testing. Holm (1979) has proposed a step-down procedure which is more liberal, but when the number of hypotheses is large, it is nearly as conservative as the Bonferroni procedure. A popular alternative is permutation resampling (Westfall and Young 1993; Ge et al. 2003). This method shuffles the phenotypes of the study subjects a large number of times in order to create permuted data sets from the null distribution of no genotype-phenotype association. Computing the test statistic for each of these permuted datasets generates the empirical joint distribution and adjusted p-values for which the actual data structure is incorporated in the multiple test correction.

While this is an improvement over the Bonferroni correction, it has limitations. Permutation is computationally demanding, since it entails repeating possibly extensive analyses many times in order to generate the empirical joint distribution. Recent developments have decreased the time (Dudbridge and Koeleman 2004) required for permutation. However, the computation of complex test statistics for many hypotheses such as is required in genomewide association studies may still be overwhelming. In addition, permutation requires complete exchangeability under the null hypothesis, so may not be applicable when there are covariates or nuisance parameters (Lin 2005).

We propose a computationally efficient method to properly adjust for multiple testing in large-scale association studies. This method can be used to control the

probability of k (≥ 1) or more false positives, denoted by k-FWER (Lehmann and Romano 2005). They derive the Bonferroni k-FWER adjustment procedure as rejection of a p-value if it is below the threshold of $k\alpha/m$, where there are m total hypotheses being tested at a nominal significance level α. The corresponding Holm stepdown procedure is to reject the ith smallest p-value if it is less than α_i, where α_i is given by: $\alpha_i = k\alpha/m$ if $i \leq k$ and $\alpha_i = k\alpha/(s + k - i)$ if $i > k$. As in the case of $k = 1$, though, both of these procedures are overly conservative.

The basic strategy of the proposed method is to ascertain the joint distribution of the test statistics among windows and to evaluate this joint distribution by an efficient Monte Carlo procedure. By properly accounting for the correlations of the test statistics, the proposed method avoids the conservativeness of the Bonferroni approach. Our Monte Carlo procedure reduces the computational burden by orders of magnitude in comparison to permutation. Simulation studies with the phased haplotypes of the Caucasian HapMap population showed that the proposed method provides accurate control of the traditional FWER as well as the more general k-FWER with various choices of window. It is substantially more powerful than the Bonferroni correction and the k-FWER version of Lehmann and Romano (2005).

We applied the new method to a case-control study of association between rheumatoid arthritis and 2,300 SNPs in a region of interest on chromosome 18. Previous studies had shown mild evidence for linkage in this region (Merriman et al. 2001) as well as possible links of this region to a variety of other auto-immune diseases such as type I diabetes and multiple sclerosis. The single-SNP analysis did not find any significant results (after adjusting for multiple comparisons), and neither did the haplotype-based analysis with the Bonferroni correction. The use of the proposed method revealed several areas that merit further investigations.

4.1 Methods

Our strategy for testing haplotype effects in case-control data can be broken up into three main steps: selecting a set of windows, assessing association within each window, and adjusting for multiple comparisons across windows. We first consider a windowing framework. In general, this will consist of all windows of S adjacent SNPs which overlap by anywhere from 0 to $S-1$ SNPs. The window size S may be chosen based on prior knowledge of a haplotype size, or exhaustive testing of a range of values for S may be performed. Within a given window, then, we first estimate the frequencies of all possible haplotypes for cases and controls separately by using the EM algorithm (Excoffier and Slatkin 1995). To improve stability and speed up computation, we remove the haplotypes with estimated frequencies $< c_f$ in the control group, where c_f is a very small number, say $1/n$ or $2/n$, and n is the total number of subjects in the study. The remaining number of haplotypes is denoted by K. Individuals whose genotypes are not compatible with the remaining set of haplotypes are dropped from the data.

As in the previous chapter, for $k = 1, \ldots, K$, let h_k denote the kth haplotype and let π_k denote the frequency of h_k in the whole population. We fit a logistic regression model with additive haplotype effects for all haplotypes with estimated frequencies $> c_e$ in both cases and controls, where c_e is a small number, say $5/n$ or $10/n$. We use the most frequent haplotype as the reference group in the model. The haplotypes with estimated frequencies less than the threshold c_e are also included in the reference group. It would be difficult to detect separate effects of such rare haplotypes. The number of haplotype effects in the model is denoted by r.

The observed data consist of (Y_i, G_i), $i = 1, \ldots, n$, where Y_i and G_i denote the disease status and genotype for the ith subject. With H representing the pair of haplotypes for a subject, the logistic regression model takes the form:

$$\Pr(Y = 1 | H = (h_k, h_l)) = \frac{e^{\alpha + \beta^T Z(h_k, h_l)}}{1 + e^{\alpha + \beta^T Z(h_k, h_l)}},$$

where α pertains to the intercept, β represents log-odd ratios,

$$Z(h_k, h_l) = \begin{bmatrix} I(h_k = h_1) + I(h_l = h_1) \\ \vdots \\ I(h_k = h_r) + I(h_l = h_r) \end{bmatrix},$$

and $I(\cdot)$ is the indicator function. The simplifications to the likelihood made in the previous chapter under the assumptions of rare disease and HWE apply to this model as well. The primary difference from the previous chapter and from Lin et al. (2005) in this setup is that we focus on an overall test for the effects of all the haplotypes, rather than tests for individual effects of haplotypes.

To incorporate the constraints that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$, $k = 1, \ldots, K$, into the calculations, we reparametrize the model by defining $\pi_k^* = \pi_k / \pi_K$ and $\nu_k = \log \pi_k^*$, $k = 1, \ldots, K$. Write $\nu = (\nu_1, \ldots, \nu_{K-1})$ and $\theta = (\beta, \nu)$. Then the log-likelihood can be written as

$$l(\theta) = \sum_{i=1}^{n} \log \left[\frac{\sum_{k,l} I\left[(h_k, h_l) \in S(G_i)\right] e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} e^{\theta^T W(Y_i, h_k, h_l)}} \right],$$

where

$$W(Y_i, h_k, h_l) = \begin{bmatrix} Y_i Z(h_k, h_l) \\ I(h_k = h_1) + I(h_l = h_1) \\ \vdots \\ I(h_k = h_{K-1}) + I(h_l = h_{K-1}) \end{bmatrix}.$$

The corresponding score function and information matrix are

$$U(\theta) = \sum_{i=1}^{n} \left[\frac{\sum_{k,l} I\left[(h_k, h_l) \in S(G_i)\right] W(Y_i, h_k, h_l) e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} I\left[(h_k, h_l) \in S(G_i)\right] e^{\theta^T W(Y_i, h_k, h_l)}} \right.$$

$$\left. - \frac{\sum_{k,l} W(Y_i, h_k, h_l) e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} e^{\theta^T W(Y_i, h_k, h_l)}} \right],$$

and

$$\Sigma(\theta) = \sum_{i=1}^{n} \left[\frac{\sum_{k,l} W(Y_i, h_k, h_l)^{\otimes 2} e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} e^{\theta^T W(Y_i, h_k, h_l)}} - \left\{ \frac{\sum_{k,l} W(Y_i, h_k, h_l) e^{\theta^T W(Y_i, h_k, h_l)}}{\sum_{k,l} e^{\theta^T W(Y_i, h_k, h_l)}} \right\}^{\otimes 2} \right]$$

$$-\sum_{i=1}^{n}\left[\frac{\sum_{k,l} I\left[(h_k,h_l)\in S\left(G_i\right)\right] W(Y_i,h_k,h_l)^{\otimes 2} e^{\theta^T W(Y_i,h_k,h_l)}}{\sum_{k,l} I\left[(h_k,h_l)\in S\left(G_i\right)\right] e^{\theta^T W(Y_i,h_k,h_l)}}\right.$$
$$\left.-\left\{\frac{\sum_{k,l} I\left[(h_k,h_l)\in S\left(G_i\right)\right] W(Y_i,h_k,h_l) e^{\theta^T W(Y_i,h_k,h_l)}}{\sum_{k,l} I\left[(h_k,h_l)\in S\left(G_i\right)\right] e^{\theta^T W(Y_i,h_k,h_l)}}\right\}^{\otimes 2}\right],$$

where $a^{\otimes 2} = aa^T$. To obtain the maximum likelihood estimate $\widehat{\theta}$, we solve the score equation $U(\theta) = 0$ by using the Newton-Raphson method. We set the initial value of θ to $\widetilde{\theta} = (0, \widetilde{\nu})$, where $\widetilde{\nu}$ is the maximum likelihood estimate of ν in the pooled sample obtained by the EM algorithm.

We can test the haplotype-disease association by using the likelihood ratio statistic $2[l(\widehat{\theta}) - l(\widetilde{\theta})]$, the score statistic, or the Wald statistic. All three test statistics have approximately the χ^2 distribution with r degrees of freedom. In deriving the joint distribution of the test statistics over different windows, it is convenient to work with the score statistic. We partition the score function and information matrix to conform with the partition of β and ν in θ, i.e.,

$$U(\theta) = \begin{bmatrix} U_\beta(\theta) \\ U_\nu(\theta) \end{bmatrix},$$

and

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{\beta\beta}(\theta) & \Sigma_{\beta\nu}(\theta) \\ \Sigma_{\nu\beta}(\theta) & \Sigma_{\nu\nu}(\theta) \end{bmatrix}.$$

Also, let $U_{\beta,i}(\theta)$ and $U_{\nu,i}(\theta)$ denote the contributions from the ith subject to $U_\beta(\theta)$ and $U_\nu(\theta)$. The score statistic can then be written as

$$T = U_\beta(\widetilde{\theta})^T V^{-1} U_\beta(\widetilde{\theta}),$$

where $V = \sum_{i=1}^n U_i U_i^T$ and $U_i = U_{\beta,i}(\widetilde{\theta}) - \Sigma_{\beta\nu}(\widetilde{\theta})\Sigma_{\nu\nu}^{-1}(\widetilde{\theta}) U_{\nu,i}(\widetilde{\theta})$.

We approximate the joint distribution of the test statistics over windows through a Monte Carlo simulation procedure. Specifically, we construct $\widetilde{T} = \widetilde{U}^T V^{-1} \widetilde{U}$, where $\widetilde{U} = \sum_{i=1}^n U_i X_i$, and X_i, $i = 1, \ldots, n$, are independent standard normal random variables. Suppose that we have a total of m windows, which may or may not be overlapping and which covers the whole region one is scanning. Let T_j and $\widetilde{T}_j$ denote

the values of T and $\widetilde{T}$ in the jth window. The same set of X_i, $i = 1, \ldots, n$, is used for all m simulated statistics $\widetilde{T}_1, \ldots, \widetilde{T}_m$. By the arguments of Lin (2005), the joint distribution of $(T_1, \ldots, T_m)$ can be approximated by the joint distribution of $(\widetilde{T}_1, \ldots, \widetilde{T}_m)$. We obtain realizations from the latter distribution by generating the normal samples $(X_1, \ldots, X_n)$ while fixing the genotype and phenotype data at their observed values.

While the simulated statistics are based on the score test, the above Monte Carlo approximation is valid whether the observed $T_1, \ldots, T_m$ are the likelihood ratio, score or Wald statistics. Our simulation studies revealed that the approximation tends to be more accurate for the likelihood-ratio statistics than the score and Wald statistics although the differences are generally very small. The numerical results reported in this article pertain to the likelihood ratio.

In the standard multiple-testing framework (Westfall and Young, 1993; Lin, 2005), the m test statistics have the same degrees of freedom. In our setting, the test statistics have different degrees of freedom because the number of haplotype effects tested varies among windows. Thus, we propose a step-down multiple-testing procedure which orders the p-values of the test statistics rather than the actual values of the test statistics. This is similar to Algorithm 2.8 in Westfall and Young (1993), which uses resampling rather than Monte Carlo methods to simulate p-values.

For $j = 1, \ldots, m$, let p_j be the (observed) p-value associated with the test statistic T_j, which is obtained from the χ^2 distribution with r_j degrees of freedom, where r_j is the number of haplotype effects tested in the jth window. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered p-values, and let $H_{(1)}, \ldots, H_{(m)}$ be the corresponding null hypotheses. In addition, let $\widetilde{T}_{(1)}, \ldots, \widetilde{T}_{(m)}$ be the simulated test statistics associated with $H_{(1)}, \ldots, H_{(m)}$, and let $\widetilde{p}_{(1)}, \ldots, \widetilde{p}_{(m)}$ be the corresponding simulated p-values. The adjusted p-value for testing $H_{(j)}$ (i.e., the smallest significance level at which $H_{(j)}$ would be rejected by the multiple testing procedure) is determined by

$$\Pr\left(\min_{j \leq l \leq m} \widetilde{p}_{(l)} \leq p_{(j)}\right).$$

We estimate this probability with 5,000 realizations of $\widetilde{p}_{(1)}, \ldots, \widetilde{p}_{(m)}$, which are obtained by repeatedly generating the normal samples $(X_1, \ldots, X_n)$ while holding the observed data fixed. To control the traditional family-wise error rate at α, one would reject only those hypotheses whose adjusted p-values are less than α.

The traditional family-wise error rate (FWER) may be too stringent in massive-scale hypothesis testing. Thus, we extend Lehmann and Romano's (2005) idea of controlling k-FWER, which is the probability of rejecting greater than or equal to k true hypotheses. To obtain the Monte Carlo adjusted p-values on the basis of k-FWER, we simply replace the minimum p-value in the above formula by the kth smallest p-value. The adjusted p-values based on the Bonferroni correction, as suggested by Lehmann and Romano, are $p_{(j)}m/k$, $j = 1, \ldots, m$. Only $k = 1$ and 2 were used in our calculations in this article, although a larger value of k may be desirable for increased quantities of markers.

4.2 Simulation Studies

We simulated data from the 120 phased haplotypes of Caucasians in the Phase I HapMap data. We considered two regions on chromosome 18: the ENCODE region, which consists of 796 SNPs, and the full set of 32,177 SNPs for the chromosome. We selected a pair of haplotypes randomly from the HapMap data for each subject and then added the two haplotypes to give the subject's genotype. We generated disease according to an additive-effect logistic model with an overall disease rate of five percent.

Figures 4.1 and 4.2 display the locations of the two HapMap regions and the linkage disequilibrium (LD) among the SNPs. The SNPs in the ENCODE region show much higher levels of LD than the full set of SNPs. This reflects the fact that the density of SNPs in the ENCODE region is higher than elsewhere.

We used the ENCODE data to assess the performance of the proposed Monte Carlo method and Bonferroni correction for different window sizes, and overlapping versus

non-overlapping windows. We set $c_f = 2/n$ and $c_e = 10/n$. We considered both the FWER and 2-FWER, denoted by Bon and Bon-2 for the Bonferroni correction and by MC and MC-2 for the proposed Monte Carlo (MC) method. The results of these studies for windows of three and four SNPs are presented in Table 4.1. For both size windows, the causative haplotype began at the 601st SNP and had frequency of .14. The type I error pertains to the probability of declaring any disease-causing SNPs when no effect exists, while the power is the same quantity when one haplotype is in fact causative. Both MC and MC-2 provide accurate control of the type I error in all cases, whereas both Bon and Bon-2 are severely conservative and thus much less powerful than MC and MC-2. As expected, MC-2 is considerably more powerful than MC. The power of MC is similar to and often higher than that of Bon-2. Using the proposed method, a sample size of 2,000 subjects is sufficient to detect an odds ratio of 1.5 with high power, and even a sample size of 1,000 provides power $> .8$ for an odds ratio of 1.7. Non-overlapping windows appear to have higher power than overlapping windows.

We can compare these results to single-SNP analysis of data generated using the same causal haplotypes, to see how much additional power one gains by looking at haplotypes relative to a single locus using the Monte Carlo procedure. For a sample size of 1000, and an odds ratio of 1.5, there is a reduction in power of about 3% for both window sizes of 3 and 4. This is not substantial, primarily due to the simulation setup, where we have a fairly common causal haplotype. The haplotype analysis will have increased improvement in power when there is a rare causative SNP which is captured by a rare haplotype, but not actually measured in the data. This has been mentioned previously (de Bakker et al. 2005), and we have also performed simulations in this vein, and, depending upon SNP frequency and level of linkage disequilibrium, seen improvements in power of up to 25% for the haplotype analysis over the single-SNP analysis (results not shown).

For the full set of 32,177 SNPs, we used non-overlapping windows of size 3, and the results are presented in Table 4.2. The causative haplotype began at the 637th

SNP, and had frequency .18. As expected, the larger magnitude of data in this setting leads to lower all around power as compared to the ENCODE data. The decline in power, however, is not drastic in view of the fact that the number of tests is increased by a factor of 40.

In the above two sets of studies, causative haplotypes had the same length as the window size used for analysis, so the power would be higher than what might be expected in a real study, where the length of the disease-predisposing haplotype is unknown. Thus, we considered exhaustive testing of non-overlapping windows of one to four SNPs in the ENCODE data. The results are presented in Table 4.3. The increase in the multiplicity of tests seems to cause only a slight loss of power in comparison to the non-overlapping windows of a fixed size.

We conducted another set of simulation studies to assess the sensitivity of our method to various assumptions. To increase genetic diversity, we generated data from the full set of SNPs on chromosome 18 according to the algorithm of Durrant et al. (2004). The causative haplotype had a frequency of .18 and was located at the same window of SNPs as in the previous studies using the full set of SNPs, which started at the 637th SNP. We generated haplotypes under the following form of Hardy-Weinberg disequilibrium:

$$\pi_{kl} = \begin{cases} \pi_k^2 + \rho\pi_k(1-\pi_k), k = l, \\ (1-\rho)\pi_k\pi_l, k \neq l, \end{cases}$$

where $\rho = .02$ (Lin et al. 2005). We increased the overall disease rate to 10% and decreased thresholds c_f and c_e to $1/n$ and $5/n$, respectively. We considered 10,000 windows of 5 SNPs, each overlapping by 3 SNPs. The results are presented in Table 4.4. The Monte Carlo method continues to have correct type I error while the Bonferroni correction remains conservative. The relative power of Bon, Bon-2, MC and MC-2 has the same trend as in the previous studies.

4.3 Rheumatoid Arthritis Dataset

Study subjects were taken from the North American Rheumatoid Arthritis Consortium (NARAC). Numerous studies (Plenge et al. 2005; Jawaheer et al. 2004) have used data from this source, and details of enrollment procedures have been previously published (Jawaheer et al. 2001). Detailed clinical and marker data are available on the NARAC website (http://www.naracdata.org). Families in this consortium satisfied the following requirements: two or more siblings fulfilled the American College of Rheumatology (ACR) 1987 criteria for rheumatoid arthritis (RA) (Arnett et al. 1998); at least one sibling had documented erosions on hand radiographs; and at least one sibling had disease onset between the ages of 18 and 60 years. Families with any other disease associated with similar articular symptoms, such as psoriasis or inflammatory bowel disease, were excluded. A total of 460 cases were chosen from throughout the United States, and confirmation of RA diagnosis was obtained from patients' rheumatologists. Radiographs of the hands and wrists were also obtained to document the presence and extent of joint involvement. A total of 460 unrelated controls from Long Island were matched to the cases on the basis of age and sex. All subjects are non-Ashkenazi Caucasians. Informed consent was obtained from all subjects, and approval of the local institutional review board was secured at every recruitment site prior to enrollment.

The SNPs were a custom set selected from dbSNP "double hit" SNPs on the basis of their distribution and favorable assay design characteristics. The 2297 SNPs represent the SNPs successfully typed with minor allele frequency greater than 5% out of the 3072 SNPs attempted in a region of chromosome 18; the region is shown in Figure 4.1 (a). The assumption of Hardy-Weinberg equilibrium (HWE) was examined for single markers using the exact test implemented in Merlin. Several were identified with significant deviations from HWE, even though neighboring markers often showed good coincidence between observed and expected genotype frequencies. Because some significant deviations from HWE are expected by chance even when the assumption

holds and departures from HWE may be caused by association between the marker alleles and disease susceptibility, we did not exclude any markers from the analysis.

We applied the proposed Monte Carlo method as well as the Bonferroni and permutation methods to this study, and considered both FWER and 2-FWER. We set $c_f = 1/n$ and $c_e = 10/n$. The results for non-overlapping windows of size 4 are summarized in Table 4.5; only the windows with adjusted MC-2 p-values of less than .25 are shown. The last two windows, D and E, in the table merit special attention, as their MC-2 p-values are less than .1. As expected, the MC and MC-2 adjusted p-values are much smaller than their Bonferroni counterparts. Indeed, the Bonferroni adjusted p-values are two to three fold of their MC counterparts. For this study, permutation was computationally feasible (although very slow) and yielded similar results to those of the MC method. Table 4.6 identifies the SNPs and the most significant haplotypes in the 5 windows with MC-2 adjusted p-values of less than .25.

There were no significant SNPs in the single-SNP analysis, whether with the simple Bonferroni correction or the more powerful MC method. The lowest adjusted p-value for any single SNP was a MC-2 p-value of 0.16. The single-SNP analysis yielded unadjusted p-values of .621, .554, .077, and .151 for the 4 SNPs in window E, which has an unadjusted p-value of .0005 for the overall haplotype test and an unadjusted p-value of .0025 for the effect of haplotype 1111. Thus, the haplotype analysis provides much stronger evidence for genetic effects than the single-SNP analysis in this study.

We also performed exhaustive testing of non-overlapping windows of sizes one through four, which did not produce any significant SNPs or windows. This is not surprising, as this procedure entails more than eight times as many tests as the analysis of non-overlapping windows of size 4, which had only mildly significant results. In this study, the gain in power from looking at different size windows did not compensate for the extra quantity of tests.

In summary, the proposed MC-2 method produced two adjusted p-values of less than .1. This degree of significance was achieved because the analysis made use of haplotypes and 2-FWER. No adjusted p-value would be less than .1 if the analysis

was based on individual SNPs, traditional FWER, or Bonferroni correction.

4.4 Discussion

The proposed method incorporates several new ideas: (1) a stable and efficient algorithm was constructed to calculate a proper statistic for testing haplotype-disease association for a given window of SNPs; (2) the joint distribution of such test statistics over different windows was derived; (3) the concept of k-FWER was adopted; (4) an accurate Monte Carlo procedure for multiple testing was developed. The concept of k-FWER is useful in genomewide association studies even if one is not interested in haplotype analysis.

Like Epstein and Satten (2003) and Lin et al. (2005), our statistic for testing haplotype-disease association for a set of SNPs is based on the retrospective likelihood, which properly reflects the case-control sampling. The calculation of our test statistic makes use of a novel parameterization, which lends itself to a simple Newton-Raphson algorithm that is more efficient and more reliable than the EM-algorithms used by the previous authors. More important, this article deals with haplotype analysis in association scans rather than candidate genes, and demonstrates improvement in power over the single SNP Monte Carlo analysis considered in Lin (2005). As noted previously, single SNP analysis may also have reduced power to detect rare causative SNPs.

Caution must be used in interpreting odds ratios from windows selected by the proposed method, given that any genomewide scan consists of selecting for the most extreme statistics. Garner (2007) discusses this point in detail for simple single SNP test statistics, and concludes that it is possible to achieve unbiased odds ratio estimates with large enough sample size. This warrants further investigation to determine the sample size necessary for more powerful methods. Meanwhile, it is important to be aware of the bias inherent in effect estimation from genomewide studies, and of the necessity for replication studies.

Our analysis of the rheumatoid arthritis study suggested loci for further investigations. Our collaborators at the North American Rheumatoid Arthritis Consortium are currently genotyping an additional 667 cases and 662 controls in the regions shown in Table 4.6. Furthermore, an independent set of cases and controls from Europe will be used for confirmation.

The selection of model parameters and window framework for the suggested procedure requires some thought. The values of c_f and c_e determine the level at which rare haplotypes are either removed completely from the dataset, or omitted from effects testing. Lower values for both thresholds permit greater characterization of rare haplotypes and their association with disease; however, overly low values will destabilize the algorithm. We have presented analyses using a range of thresholds which perform well in practice. An alternative is to incorporate the haplotype clustering methods of Tzeng et al. (2006) into the test for haplotype-disease association within a window. As these methods are formulated in terms of the score test of Schaid et al. (2002), multiplied by an allocation matrix, they could be used within the Monte Carlo framework described here.

The number of windows to be used must be balanced against the degree of penalty for multiple testing. It may be more powerful to focus on non-overlapping windows than to consider every possible adjacent group of SNPs. One compromise is to use exhaustive testing with non-overlapping windows. The level of LD in the region of causative SNPs will also affect the power. A lower level of LD will create a greater number of common haplotypes, and thus will reduce the power to detect a true effect. In regions of high LD, non-overlapping windows will certainly have high power even if the causative haplotype happens to be out of phase with the windows. A longer causative haplotype will not be as well detected by windows of 4 or 5 SNPs as by a larger window. Testing for larger windows increases the computational intensity greatly, because of the increase in the numbers of haplotypes. The need for testing with large windows can be alleviated by using tag SNPs. If a few SNPs encode much of the variation in a region, then a small set of tag SNPs can capture the effect of a

long haplotype.

We have selected windows without considering the actual LD patterns. An alternative approach is to select windows in such a way that the SNPs are in strong LD within windows and in low LD between windows. A simple approach would be to select non-overlapping windows based on a definition of haplotype blocks, such as all SNPs within the block having pairwise correlation $> .8$. This allows for variable length blocks in analysis; however, it is not without its own problems, such as the somewhat arbitrary definition of haplotype blocks. Li et al. (2007) implement an alternate procedure in which the maximum size of a sliding window is determined by local haplotype diversity and sample size. It would be worthwhile to investigate the performance of such strategies.

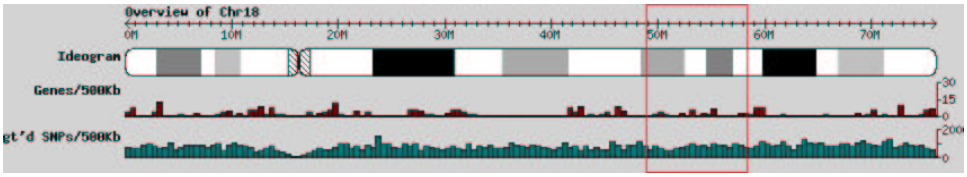
Finally, the choice of k for controlling the k-FWER will affect the interpretation of results. It is clear that using the k-FWER results in higher power, as a consequence of relaxing the significance threshold. However, this increase in power is accompanied by an increased expected number of false positives. An alternative would be to increase the alpha level for $k = 1$. For Bonferroni, these two procedures are equivalent, since doubling the alpha level has the same effect as computing the 2-FWER adjusted p-values. In general, though, it is not clear what significance threshold for $k = 1$ is equivalent to controlling the 2-FWER at the .05 level. Chen and Storey (2006) discuss a similar measure, GWER-k, for linkage analysis, where the GWER-k is equivalent to the $(k + 1)$-FWER. In their simulations, controlling the GWER-1 at the .05 level resulted in GWER-0 rates which ranged from .13 to .34. Considering different values of k may thus be more practical than attempting to achieve the same increase in power by increasing the threshold for $k = 1$.

This chapter is focused on genetic effects. In some studies, investigators are interested in gene-environment interactions. By incorporating the profile likelihood approach of Lin et al. (2005), we can extend the proposed method to detect haplotype-environment interactions in genomewide association studies. In addition, we may accommodate Hardy-Weinberg disequilibrium as in Lin et al. (2005).

The proposed Monte Carlo procedure is substantially more powerful than the conventional Bonferroni correction while providing accurate control of the type I error. The Monte Carlo procedure requires nearly a thousandth the computing time of the permutation procedure (with 1,000 permuted data sets) and thus can be used for studies involving large quantities of SNPs. This differential is due to the fact that the time necessary to generate the simulated statistics is negligible compared to that necessary to calculate the observed test statistics. Hence permutation, which performs the latter procedure 1000 times, is much more time consuming than the Monte Carlo. For the rheumatoid arthritis study, it took about 320 seconds on an IBM BladeCenter HS20 machine to carry out the Monte Carlo procedure for non-overlapping windows of 4 SNPs, as opposed to 39 hours for permutation. Exhaustive testing for windows ranging from 1 SNP to 4 SNPs required 1225 seconds. It would be more difficult to use permutation if one is interested in testing gene-environment interactions.

Lin et al. (2004) considered exhaustive testing of haplotype-disease association over all possible windows of segments, and used a computationally efficient permutation procedure to assess the significance of the correlated tests. Their approach is based on a version of the transmission disequilibrium test and is applicable to family data only. Our approach can also be extended to family studies.

a.



b.

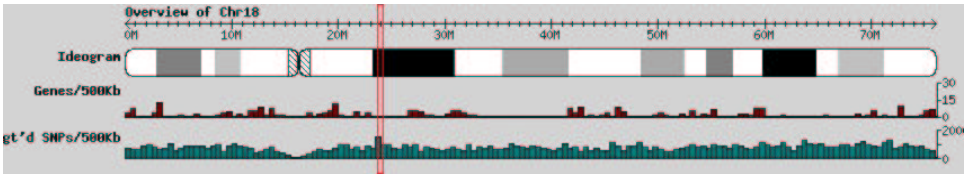
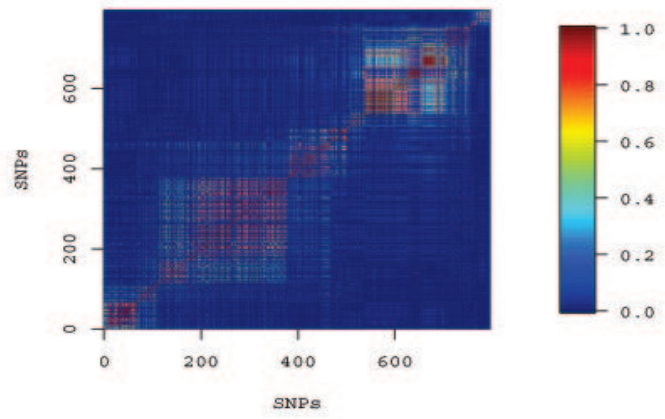


Figure 4.1: Locations of SNPs in two regions of interest on chromosome 18: (a) 2300 SNPs from the rheumatoid arthritis case-control study; (b) 796 SNPs from the HapMap ENCODE region

a.



b.

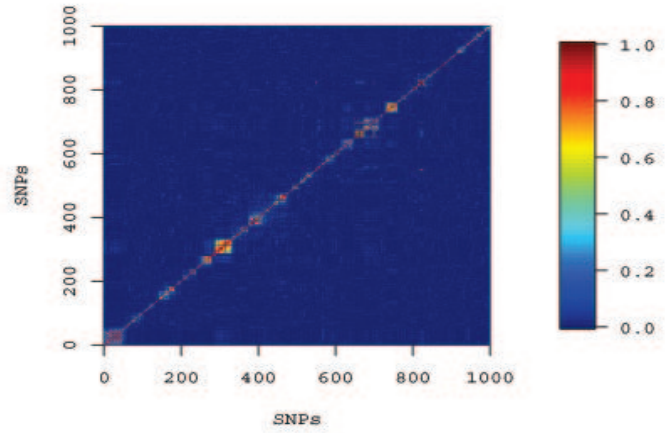


Figure 4.2: Patterns of LD, as measured by the squared correlation coefficient r^2 between pairs of markers, in two HapMap regions on chromosome 18: (a) 796 SNPs in the ENCODE region; (b) first 1000 SNPs in the full set of SNPs

Table 4.1: Type I error/power of haplotype tests at the .05 nominal significance level based on the ENCODE data.

Odds	Sample	Overlapping windows				Non-overlapping windows			
ratio	size	Bon	Bon-2	MC	MC-2	Bon	Bon-2	MC	MC-2
				Windows of 3 SNPS					
1.0	1000	.014	.016	.041	.041	.015	.017	.054	.055
	2000	.009	.015	.041	.049	.016	.010	.047	.039
1.5	1000	.369	.455	.488	.557	.371	.447	.509	.611
	2000	.754	.798	.880	.914	.798	.853	.876	.931
1.7	1000	.693	.741	.843	.899	.725	.786	.829	.900
				Windows of 4 SNPS					
1.0	1000	.007	.009	.038	.043	.023	.018	.046	.040
	2000	.013	.013	.049	.053	.017	.009	.040	.036
1.5	1000	.286	.356	.481	.572	.397	.480	.512	.631
	2000	.735	.791	.876	.907	.813	.863	.879	.935
1.7	1000	.676	.736	.821	.867	.743	.808	.845	.907

Note: The total sample size is given; there are equal numbers of cases and controls. Bon and Bon-2 pertain to the FWER and 2-FWER based on the Bonferroni correction, and MC and MC-2 to the FWER and 2-FWER based on the Monte Carlo procedure. Each entry is based on 1,000 simulated datasets.

Table 4.2: Type I error/power of haplotype tests with non-overlapping windows of 3 SNPs at the .05 nominal significance level based on the full set of SNPs on chromosome 18 of the HapMap data for studies with 500 cases and 500 controls.

Odds ratio	Bon	Bon-2	MC	MC-2
1.0	.022	.015	.039	.046
1.5	.221	.296	.292	.477
1.7	.577	.649	.646	.761

Note: Bon and Bon-2 pertain to the FWER and 2-FWER based on the Bonferroni correction, and MC and MC-2 to the FWER and 2-FWER based on the Monte Carlo procedure. Each entry is based on 1,000 simulated datasets.

Table 4.3: Type I error/power of the exhaustive testing with non-overlapping windows of 1-4 SNPs based on the ENCODE data when the causative haplotype contains 4 SNPs.

Odds ratio	Sample size	Bon	Bon-2	MC	MC-2
1.0	1000	.016	.010	.051	.057
	2000	.011	.016	.057	.060
1.5	1000	.308	.388	.480	.567
	2000	.755	.818	.877	.920
1.7	1000	.677	.747	.822	.890

Note: The total sample size is given; there are equal numbers of cases and controls. Bon and Bon-2 pertain to the FWER and 2-FWER based on the Bonferroni correction, and MC and MC-2 to the FWER and 2-FWER based on the Monte Carlo procedure. Each entry is based on 1,000 simulated datasets.

Table 4.4: Type I error/power of haplotype tests with partially overlapping windows of 5 SNPs at the .05 nominal significance level under Hardy-Weinberg disequilibrium and common disease based on the full set of SNPs on chromosome 18 of the HapMap data for studies with 500 cases and 500 controls.

Odds ratio	Bon	Bon-2	MC	MC-2
1.0	.021	.016	.035	.050
1.5	.108	.157	.137	.271
1.7	.434	.492	.473	.628

Note: Bon and Bon-2 pertain to the FWER and 2-FWER based on the Bonferroni correction, and MC and MC-2 to the FWER and 2-FWER based on the Monte Carlo procedure. Each entry is based on 1,000 simulated datasets.

Table 4.5: The adjusted p-values for the 5 most significant non-overlapping windows of 4 SNPs in the rheumatoid arthritis study.

Window	Bon	Bon-2	MC	MC-2	Perm	Perm-2
A	.694	.347	.334	.137	.341	.145
B	1.000	.580	.470	.225	.479	.242
C	1.000	.553	.455	.217	.465	.234
D	.467	.234	.248	.087	.262	.090
E	.289	.144	.163	.049	.162	.047

Note: Bon and Bon-2 pertain to the FWER and 2-FWER based on the Bonferroni correction, MC and MC-2 pertain to the FWER and 2-FWER based on the Monte Carlo procedure, and Perm and Perm-2 pertain to the FWER and 2-FWER based on permutation.

Table 4.6: Estimated haplotype effects for the 5 most significant non-overlapping windows of 4 SNPs in the rheumatoid arthritis study.

Window	SNPs	Haplotype	Frequency	Odds ratio	Unadjusted p-value
A	$(377, 378, 379, 380)$	0110	.052	.46	.00052
B	$(685, 686, 687, 688)$	0110	.032	1.94	.021
		1011	.147	.70	.0096
C	$(1097, 1098, 1099, 1100)$	0110	.280	1.43	.0006
D	$(1101, 1102, 1103, 1104)$	0100	.305	1.41	.00083
E	$(1141, 1142, 1143, 1144)$	1111	.030	2.54	.0025
		0001	.053	.67	.061

Chapter 5

Efficient Association Mapping of Quantitative Trait Loci with Selective Genotyping

Case-control studies are a popular design, particularly for genomewide association studies. However, mapping genes associated with quantitative traits is an important step toward genetic dissection of complex human diseases. Disease genes are unlikely to have very large effects on quantitative traits, so power is a major concern in association studies, especially with the need to adjust for multiple testing. Despite the continuing improvements in genotyping efficiency, it is still highly expensive to genotype a large number of individuals, particularly in genomewide association studies. A cost-effective strategy is to preferentially genotype individuals whose trait values deviate from the population mean. Known as selective genotyping, this approach can result in a substantial increase in power (relative to random sampling with the same number of individuals) because much of the genetic information resides in individuals with extreme phenotypes (Laitinen et al. 1997; Slatkin 1999; van Gestel et al. 2000; Xiong et al. 2002; Chen et al. 2005; Cornish et al. 2005; Wallace et al. 2006).

5.1 Selective Genotyping and Outcome Dependent Sampling

Selective genotyping designs are a subclass of outcome-dependent sampling (ODS). Such designs can be defined as retrospective sampling schemes where one observes the exposure/covariates with a probability which depends on the outcome variable. The main idea is to concentrate resources where there is the greatest amount of information (Zhou et al. 2002). The case-control design is a well-known example, where cases are sampled with a greater frequency than in the general population in order to increase the available information in the sample.

The retrospective nature of ODS designs adds complexity to analysis. Prentice and Pyke (1979) showed that the prospective analysis ignoring selection, and retrospective analysis, result in the same estimators for the logistic regression in case-control studies. While this is true in some situations (Chen 2003), it does not hold for general ODS designs. Indeed, we compare the prospective analysis with proper maximum likelihood estimation under two selective genotyping designs, and show that the prospective analysis is biased and has reduced power.

Recently, semiparametric methods have been developed for several different ODS designs (Zhou et al. 2002; Lawless et al. 1999). Lawless et al. (1999) consider the situation where the observation of trait and covariates depends on which of a finite number K of strata the trait belongs to. They develop semi-parametric maximum likelihood and pseudolikelihood estimation for two-phase designs. Zhou et al. (2002) consider a semiparametric empirical likelihood inference procedure in which the underlying distribution of covariates is treated as a nuisance parameter and is left unspecified. The ODS design in this study includes a simple random sample from the population, supplemented by samples drawn from particular regions of the outcome space. While these designs are useful for a broad class of designs, the methods developed are not applicable in haplotype-based selective genotyping designs, since haplotypes are not directly observed.

Selective genotyping has recently developed into a class of designs distinct from general ODS designs due to the special nature of haplotypes and genotypes. Slatkin (1999) suggested genotyping a selected sample of individuals with unusually high values of the quantitative trait, together with a random sample from the study population. Because selection depends on the phenotype, standard statistical methods (e.g., t-test and ANOVA), which assume random sampling, are inappropriate. Slatkin (1999) developed two tests: one comparing the allele frequencies between the selected sample and random sample, and one comparing the mean trait values among individuals with different genotypes in the selected sample. The two tests are approximately independent, so their p-values can be combined to form an overall test. Slatkin (1999) used simulation to show that his tests are more powerful than the simple t-test (when the latter is applied to a random sample with the same number of individuals). Chen et al. (2005) recommended replacement of the random sample with a selected sample of individuals with unusually low trait values and described two sampling schemes to obtain the selected samples. They demonstrated through a simulation study that, using Slatkin's three tests, their designs are more efficient than Slatkin's original design.

In a recent *Science* report on obesity (Herbert et al. 2006), one of the replication studies genotyped individuals from the 90th to 97th percentile of the BMI distribution and those from the 5th to 12th percentile, and another replication study genotyped individuals from the top and bottom quartiles. In both studies, the individuals with high and low BMI values were treated as cases and controls, respectively, and case-control methods (i.e., testing for allele-frequency differences between the two selected groups) were used for analysis.

Case-control methods disregard the actual trait values and are thus inefficient. Slatkin's (1999) tests do not make full use of the available data either — individuals who are homozygous for the minor allele are discarded, and the trait values in the random sample or the low-trait-value sample are not used at all. Recently, Wallace et al. (2006) proposed a Hotelling's T^2 test for normal traits, which they showed through simulation has increased power over Slatkin's tests. Wallace et al.'s (2006)

test, which is essentially the standard t-test in the case of a single marker, ignores the biased sampling nature of the selective-genotyping design and thus may be inefficient. Furthermore, none of the existing methods deal with haplotype-based testing or estimation of genetic effects.

In this chapter, we show how to properly and efficiently map quantitative trait loci (QTLs) with selective genotyping. We derive appropriate likelihoods which make full use of the available data and which properly reflect trait-dependent sampling. The corresponding inference procedures are valid and efficient. Our methods can be used to perform both genotype-based and haplotype-based association analyses.

5.2 Designs and Likelihoods

We consider two very general selective-genotyping designs. Under design 1, the quantitative trait is measured on a random sample of N individuals from the study population, and a subset of n individuals is selected for genotyping; the selection probabilities depend on the trait values. Under design 2, a random sample of n individuals whose trait values fall into certain regions are selected for genotyping, and the trait values are retained only on those individuals. Thus, the main difference between the two designs is that the trait values on those individuals who are not selected for genotyping are retained under design 1, but not under design 2. Under design 2, it is not necessary to specify N or to ascertain the individuals outside the selection regions.

Let Y_i be the trait value of the ith individual and G_i be the corresponding multi-locus genotype denoting the number of minor alleles at each SNP site. The association between G_i and Y_i is characterized by the conditional density function $P(Y_i|G_i;\theta)$ indexed by a set of parameters θ. In the special case of a single locus under the additive mode of inheritance, $P(Y_i|G_i;\theta)$ may take the familiar form of the linear regression model

$$Y_i = \alpha + \beta G_i + \epsilon_i, \tag{5.1}$$

where ϵ_i is zero-mean normal with variance σ^2. In this case, $\theta = (\alpha, \beta, \sigma^2)$. Under the dominant (or recessive) mode of inheritance, G_i in (5.1) is replaced by the indicator of whether or not the ith individual has at least one minor allele (or, for the recessive model, two minor alleles). If there are multiple loci, then βG_i in (5.1) is replaced by an appropriate linear combination of individual genotype scores and (possibly) their cross-products. We denote the probability function of the genotype by $P(G;\gamma)$, where γ represents the (multi-locus) genotype frequencies.

Under design 1, the data consist of (Y_i, G_i) $(i = 1, \ldots, n)$ and Y_i $(i = n+1, \ldots, N)$. (Without loss of generality, the data are so arranged that the first n records pertain to the n individuals who are selected for genotyping and the remaining $(N-n)$ records to the unselected individuals.) The data for design 1 can be written as $(Y_i, R_i, R_i G_i)$ $(i = 1, \ldots, N)$, where R_i indicates, by the values 1 versus 0, whether the ith individual is selected for genotyping. The likelihood function $\prod_{i=1}^{N} P(Y_i, R_i, R_i G_i)$ can be expressed as $\prod_{i=1}^{N} P(Y_i, R_i) P(R_i G_i | Y_i, R_i)$ or $\prod_{i=1}^{N} P(Y_i) P(R_i|Y_i) P(G_i|Y_i)^{R_i}$, which is proportional to $\prod_{i=1}^{N} P(Y_i, G_i)^{R_i} P(Y_i)^{1-R_i}$, because the selection probabilities $P(R_i|Y_i)$ are constants. Thus we can write the likelihood for θ and γ which corresponds to this design as

$$\prod_{i=1}^{n} P(Y_i|G_i;\theta) P(G_i;\gamma) \prod_{i=n+1}^{N} \sum_{G} P(Y_i|G;\theta) P(G;\gamma), \tag{5.2}$$

where the summation over G is taken over all possible genotypes.

Under design 2, the data consist only of (Y_i, G_i) $(i = 1, \ldots, n)$, which are a random sample from all the individuals whose trait values belong to a particular set $\mathcal{C}$. We can use the likelihood for θ and γ

$$\prod_{i=1}^{n} P(Y_i, G_i | Y_i \in \mathcal{C}) = \prod_{i=1}^{n} \frac{P(Y_i|G_i;\theta) P(G_i;\gamma)}{\sum_G P(Y_i \in \mathcal{C}|G;\theta) P(G;\gamma)} \tag{5.3}$$

or the likelihood for θ

$$\prod_{i=1}^{n} P(Y_i|G_i, Y_i \in \mathcal{C}) = \prod_{i=1}^{n} \frac{P(Y_i|G_i;\theta)}{P(Y_i \in \mathcal{C}|G_i;\theta)}. \tag{5.4}$$

If only the individuals whose trait values are less than the lower threshold c_L or larger

than the upper threshold c_U are selected for genotyping, then under model (5.1),

$$P(Y_i \in \mathcal{C}|G_i;\theta) = 1 - \Phi\left(\frac{c_U - \alpha - \beta G_i}{\sigma}\right) + \Phi\left(\frac{c_L - \alpha - \beta G_i}{\sigma}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution.

We refer to (5.2) as the full likelihood and (5.3) and (5.4) as the conditional likelihoods. These likelihoods properly reflect the selective-genotyping designs and use all the available data. Under design 1, one may disregard the trait values of those individuals who are not selected for genotyping and use the conditional likelihoods provided that the genotyped individuals are a random sample from the set $\mathcal{C}$. The maximum likelihood estimators can be obtained by the usual Newton-Raphson algorithm. By likelihood theory, the maximum likelihood estimators are approximately unbiased, normally distributed and statistically efficient. Association testing can be performed by using the familiar likelihood-ratio, score, or Wald statistics.

To show that the maximizations of (5.3) and (5.4) yield the same estimator of θ, it suffices to show that the profile likelihood for θ – that is, the maximum of expression (5.3) over γ for fixed θ – is equivalent to equation (5.4). By defining $\gamma_g = P(G = g;\gamma), n_g = \sum_{i=1}^{n} I(G_i = g)$, and $P_g(\theta) = P(Y_i \in \mathcal{C}|G = g;\theta)$, we can write the logarithm of expression (3) as $\sum_{i=1}^{n} \log P(Y_i|G_i;\theta) + \sum_g n_g \log \gamma_g - n \log \sum_g \gamma_g P_g(\theta)$. It then follows from simple algebraic manipulations that the profile log-likelihood for θ is $\sum_{i=1}^{n} \log P(Y_i|G_i;\theta) - \sum_g n_g \log P_g(\theta) + \sum_g n_g \log(n_g/n)$, which is exactly the logarithm of expression (5.4) up to the constant $\sum_g n_g \log(n_g/n)$.

The above description pertains to the analysis of genotype-phenotype association. It is also desirable to assess haplotype-phenotype association (Schaid et al. 2002; Lin et al. 2005). Let H_i denote the diplotype of the ith individual. The effects of haplotypes on the trait are characterized by the conditional density function $P(Y_i|H_i;\theta)$ indexed by a set of parameters θ. If we are interested in assessing the effect of a particular haplotype h^*, then $P(Y_i|H_i;\theta)$ may take the following form

$$Y_i = \alpha + \beta Z(H_i) + \epsilon_i, \tag{5.5}$$

where $Z(H_i)$ is the number of occurrences of h^* in H_i under the additive mode of

inheritance, the indicator of whether or not H_i contains at least one h^* under the dominant mode of inheritance, and the indicator of whether or not H_i contains two copies of h^* under the recessive mode of inheritance. One may also define $P(Y_i|H_i;\theta)$ in such a way that multiple haplotypes are compared to a reference in a single model (Lin et al. 2005).

Because haplotypes are not directly observed, it is necessary to impose some restrictions, such as Hardy-Weinberg equilibrium (HWE), on the diplotype distribution. For $k = 1, \ldots, K$, let h_k denote the kth possible haplotype in the population and let π_k denote the population frequency of h_k. Under HWE, $P(H_i = (h_k, h_l)) = \pi_k \pi_l$ $(k, l = 1, \ldots, K)$. We denote the diplotype probability function by $P(H_i;\gamma)$, where $\gamma = (\pi_1, \ldots, \pi_K)$.

Inference on haplotype effects must properly account for phase ambiguity. Note that $P(Y_i, G_i) = \sum_{H \in \mathcal{S}(G_i)} P(Y_i|H;\theta)P(H;\gamma)$, where $\mathcal{S}(G_i)$ is the set of diplotypes compatible with the observed genotype G_i (Lin et al. 2005). Thus, the full likelihood and conditional likelihood analogous to (5.2) and (5.3) are

$$\prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} P(Y_i|H;\theta)P(H;\gamma) \prod_{i=n+1}^{N} \sum_{H} P(Y_i|H;\theta)P(H;\gamma) \tag{5.6}$$

and

$$\prod_{i=1}^{n} \frac{\sum_{H \in \mathcal{S}(G_i)} P(Y_i|H;\theta)P(H;\gamma)}{\sum_{H} P(Y_i \in \mathcal{C}|H;\theta)P(H;\gamma)}, \tag{5.7}$$

where the second summation in (5.6) and the summation in the denominator of (5.7) are taken over all possible diplotypes. The maximizations of (5.6) and (5.7) can be carried out by the EM algorithm or the Newton-Raphson algorithm presented in the next sections. The maximum likelihood estimators are approximately unbiased, normally distributed and statistically efficient.

Note that β pertains to genetic effect in (5.1) and to haplotype effect in (5.5). If we are concerned with one SNP at a time, then models (5.1) and (5.5) are the same. In that case, likelihoods (5.6) and (5.7) differ from (5.2) and (5.3) in that the former impose HWE and allow missing genotype values whereas the latter do not impose

HWE and exclude subjects with missing genotype values. Thus, the former yield more efficient analyses provided that HWE is a reasonable assumption.

5.3 EM Algorithm to Maximize (5.6)

We present an EM algorithm for the maximization of (5.6) by treating the H_i as missing data. The complete-data log-likelihood is

$$\sum_{i=1}^{N}\sum_{k,l} I\{H_i = (h_k, h_l)\}\{\log P(Y_i|(h_k, h_l); \theta) + \log P((h_k, h_l); \gamma)\},$$

where $I(\cdot)$ is the indicator function. Define $p_{ikl} = P(H_i = (h_k, h_l)|Y_i, G_i)$, where G_i is unknown for $i = n+1, \ldots, N$. Then

$$p_{ikl} = \frac{I\{(h_k, h_l) \in \mathcal{S}(G_i)\}P(Y_i|(h_k, h_l); \theta)P((h_k, h_l); \gamma)}{\sum_{k,l} I\{(h_k, h_l) \in \mathcal{S}(G_i)\}P(Y_i|(h_k, h_l); \theta)P((h_k, h_l); \gamma)},$$

where $S(G_i)$ is the set of all possible diplotypes when G_i is unknown. In the E-step of the EM algorithm, we evaluate the p_{ikl} at the current estimates of θ and γ. In the M-step, we solve the following equations for θ and γ:

$$\sum_{i=1}^{N}\sum_{k,l} I\{(h_k, h_l) \in \mathcal{S}(G_i)\}p_{ikl}\partial \log P(Y_i|(h_k, h_l); \theta)/\partial\theta = 0,$$

$$\sum_{i=1}^{N}\sum_{k,l} I\{(h_k, h_l) \in \mathcal{S}(G_i)\}p_{ikl}\partial \log P((h_k, h_l); \gamma)/\partial\gamma = 0.$$

The linear regression model specifies that, conditional on $H_i = (h_k, h_l)$, the quantitative trait Y_i is normally distributed with mean $\beta^T Z(h_k, h_l)$ and variance σ^2, where $Z(h_k, h_l)$ is a specific function of h_k and h_l, and β is the corresponding set of regression parameters. If we are interested in comparing a particular haplotype h^* to all others, then $Z(h_k, h_l) = [1, I(h_k = h^*) + I(h_l = h^*)]^T$ under the additive model, $Z(h_k, h_l) = [1, I(h_k = h^*) + I(h_l = h^*) - I(h_k = h_l = h^*)]^T$ under the dominant model, and $Z(h_k, h_l) = [1, I(h_k = h_l = h^*)]^T$ under the recessive model. In this case,

$$p_{ikl} = \frac{I\{(h_k, h_l) \in \mathcal{S}(G_i)\}\exp[-\{Y_i - \beta^T Z(h_k, h_l)\}^2/(2\sigma^2)]\pi_k\pi_l}{\sum_{k,l} I\{(h_k, h_l) \in \mathcal{S}(G_i)\}\exp[-\{Y_i - \beta^T Z(h_k, h_l)\}^2/(2\sigma^2)]\pi_k\pi_l},$$

and the M-step has explicit solutions

$$\beta = \left\{ \sum_{i=1}^{N} \sum_{k,l} p_{ikl} Z(h_k, h_l) Z(h_k, h_l)^T \right\}^{-1} \left\{ \sum_{i=1}^{N} Y_i \sum_{k,l} p_{ikl} Z(h_k, h_l) \right\},$$

$$\sigma^2 = N^{-1} \sum_{i=1}^{N} \sum_{k,l} p_{ikl} \{Y_i - \beta^T Z(h_k, h_l)\}^2,$$

$$\pi_k = N^{-1} \sum_{i=1}^{N} \sum_{l=1}^{K} p_{ikl}.$$

5.4 Newton-Raphson Algorithm to Maximize (5.7)

Under the linear regression model with thresholds c_L and c_U, (5.7) becomes

$$\prod_{i=1}^{n} \frac{\sum_{(h_k,h_l)\in S(G_i)} (2\pi\sigma^2)^{-1/2} \exp\left\{-(Y_i - \beta^T Z(h_k, h_l))^2/(2\sigma^2)\right\} \pi_k \pi_l}{\sum_{k,l} \left[1 - \Phi\left(\frac{c_U - \beta^T Z(h_k,h_l)}{\sigma}\right) + \Phi\left(\frac{c_L - \beta^T Z(h_k,h_l)}{\sigma}\right)\right] \pi_k \pi_l}.$$

To incorporate the constraints that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ $(k = 1, \ldots, K)$ into the calculations, we define $\pi_k^* = \pi_k/\pi_K$ and $\eta_k = \log \pi_k^*$. For notational convenience, denote σ^2 as v. Let $\eta = (\eta_1, \ldots, \eta_{K-1})$ and $\vartheta = (\beta, v, \eta)$. Then the log-likelihood is

$$\ell(\vartheta) = -\frac{n}{2} \log v + \sum_{i=1}^{n} \log \sum_{(h_k,h_l)\in S(G_i)} \exp\left\{-(2v)^{-1}(Y_i - \beta^T Z(h_k, h_l))^2 + \eta^T W(h_k, h_l)\right\}$$

$$- n \log \sum_{k,l} e^{\eta^T W(h_k,h_l)} \left\{1 - \Phi\left(\frac{c_U - \beta^T Z(h_k, h_l)}{\sqrt{v}}\right) + \Phi\left(\frac{c_L - \beta^T Z(h_k, h_l)}{\sqrt{v}}\right)\right\},$$

where

$$W(h_k, h_l) = \begin{bmatrix} I(h_k = h_1) + I(h_l = h_1) \\ \vdots \\ I(h_k = h_{K-1}) + I(h_l = h_{K-1}) \end{bmatrix}.$$

Let

$$Q_{ikl}(\vartheta) = \exp\left\{-(Y_i - \beta^T Z(h_k, h_l))^2/(2v) + \eta^T W(h_k, h_l)\right\},$$

$$R_{kl}^L(\vartheta) = \{c_L - \beta^T Z(h_k, h_l)\}/\sqrt{v},$$

$$R_{kl}^U(\vartheta) = \{c_U - \beta^T Z(h_k, h_l)\}/\sqrt{v},$$

$$S(\vartheta) = \sum_{k,l} \left\{1 - \Phi\left(R_{kl}^U(\vartheta)\right) + \Phi\left(R_{kl}^L(\vartheta)\right)\right\} e^{\eta^T W(h_k,h_l)}.$$

Also, let $a^{\otimes 2} = aa^T$ and let ϕ be the standard normal density function. Then

$$\frac{\partial \ell(\vartheta)}{\partial v} = -\frac{n}{2v} + \sum_{i=1}^{n} \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}$$
$$- \frac{n \sum_{k,l} \left\{ \phi\left(R_{kl}^U(\vartheta)\right) R_{kl}^U(\vartheta) - \phi\left(R_{kl}^L(\vartheta)\right) R_{kl}^L(\vartheta) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)}$$

$$\frac{\partial \ell(\vartheta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\sum_{(h_k,h_l)\in S(G_i)} \frac{Q_{ikl}(\vartheta)}{v} (Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}$$
$$- \frac{n \sum_{k,l} \left\{ \phi\left(R_{kl}^U(\vartheta)\right) - \phi\left(R_{kl}^L(\vartheta)\right) \right\} \frac{Z(h_k,h_l)}{\sqrt{v}} e^{\eta^T W(h_k,h_l)}}{S(\vartheta)}$$

$$\frac{\partial \ell(\vartheta)}{\partial \eta} = \sum_{i=1}^{n} \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}$$
$$- \frac{n \sum_{k,l} \left\{ 1 - \Phi\left(R_{kl}^U(\vartheta)\right) + \Phi\left(R_{kl}^L(\vartheta)\right) \right\} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)}{S(\vartheta)}$$

$$\frac{\partial^2 \ell(\vartheta)}{\partial v^2} = \frac{n}{2v^2} + \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^4}{4v^4} - \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{v^3} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}^2 \right]$$
$$- n \left[\frac{\sum_{k,l} \left\{ \phi(R_{kl}^U(\vartheta)) \left(\left(R_{kl}^U(\vartheta)^3\right) - 3R_{kl}^U(\vartheta) \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{4v^2}}{S(\vartheta)} \right.$$
$$- \frac{\sum_{k,l} \left\{ \phi(R_{kl}^L(\vartheta)) \left(\left(R_{kl}^L(\vartheta)^3\right) - 3R_{kl}^L(\vartheta) \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{4v^2}}{S(\vartheta)}$$
$$\left. - \left\{ \frac{\sum_{k,l} \left(\phi(R_{kl}^U(\vartheta)) R_{kl}^U(\vartheta) - \phi(R_{kl}^L(\vartheta)) R_{kl}^L(\vartheta) \right) \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)} \right\}^2 \right]$$

$$\frac{\partial^2 \ell(\vartheta)}{\partial v \partial \beta} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) Z(h_k,h_l) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^3}{2v^3} - \frac{(Y_i - \beta^T Z(h_k,h_l))}{v^2} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$- \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \right]$$
$$- n \left[\frac{\sum_{k,l} \left\{ \phi(R^U_{kl}(\vartheta)) \left(R^U_{kl}(\vartheta)^2 - 1 \right) - \phi(R^L_{kl}(\vartheta)) \left(R^L_{kl}(\vartheta)^2 - 1 \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v^{3/2}} Z(h_k,h_l)}{S(\vartheta)} \right.$$
$$- \left\{ \frac{\sum_{k,l} \left(\phi(R^U_{kl}(\vartheta)) R^U_{kl}(\vartheta) - \phi(R^L_{kl}(\vartheta)) R^L_{kl}(\vartheta) \right) \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{k,l} \left(\phi(R^U_{kl}(\vartheta)) - \phi(R^L_{kl}(\vartheta)) \right) e^{\eta^T W(h_k,h_l)} \frac{Z(h_k,h_l)}{\sqrt{v}}}{S(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell(\vartheta)}{\partial (\beta\beta^T)} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{v^2} - v^{-1} \right\} Z^{\otimes 2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}^{\otimes 2} \right]$$
$$- n \left[\frac{\sum_{k,l} \left\{ \phi(R^U_{kl}(\vartheta)) R^U_{kl}(\vartheta) - \phi(R^L_{kl}(\vartheta)) R^L_{kl}(\vartheta) \right\} e^{\eta^T W(h_k,h_l)} \left(\frac{Z(h_k,h_l)}{\sqrt{v}} \right)^{\otimes 2}}{S(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{k,l} \left(\phi(R^U_{kl}(\vartheta)) - \phi(R^L_{kl}(\vartheta)) \right) e^{\eta^T W(h_k,h_l)} \frac{Z(h_k,h_l)}{\sqrt{v}}}{S(\vartheta)} \right\}^{\otimes 2} \right]$$

$$\frac{\partial^2 \ell(\vartheta)}{\partial v \partial \eta} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2} Q_{ikl}(\vartheta)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \right]$$
$$- n \left[\frac{\sum_{k,l} \left\{ \phi\left(R_{kl}^U(\vartheta)\right) R_{kl}^U(\vartheta) - \phi\left(R_{kl}^L(\vartheta)\right) R_{kl}^L(\vartheta) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v} W(h_k,h_l)}{S(\vartheta)} \right.$$
$$- \left\{ \frac{\sum_{k,l} \left\{ \phi\left(R_{kl}^U(\vartheta)\right) R_{kl}^U(\vartheta) - \phi\left(R_{kl}^L(\vartheta)\right) R_{kl}^L(\vartheta) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{k,l} \left\{ 1 - \Phi\left(R_{kl}^U(\vartheta)\right) + \Phi\left(R_{kl}^L(\vartheta)\right) \right\} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)}{S(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell(\vartheta)}{\partial \beta \partial \eta^T} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v} W(h_k,h_l)^T}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}^T \right]$$
$$- n \left[\frac{\sum_{k,l} \left\{ \phi\left(R_{kl}^U(\vartheta)\right) - \phi\left(R_{kl}^L(\vartheta)\right) \right\} \frac{Z(h_k,h_l)}{\sqrt{v}} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)^T}{S(\vartheta)} \right.$$
$$- \left\{ \frac{\sum_{k,l} \left\{ \phi\left(R_{kl}^U(\vartheta)\right) - \phi\left(R_{kl}^L(\vartheta)\right) \right\} \frac{Z(h_k,h_l)}{\sqrt{v}} e^{\eta^T W(h_k,h_l)}}{S(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{k,l} \left\{ 1 - \Phi\left(R_{kl}^U(\vartheta)\right) + \Phi\left(R_{kl}^L(\vartheta)\right) \right\} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)}{S(\vartheta)} \right\}^T \right]$$

$$\frac{\partial^2 \ell(\vartheta)}{\partial (\eta \eta^T)} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l)^{\otimes 2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}^{\otimes 2} \right]$$
$$- n \left[\frac{\sum_{k,l} \left\{ 1 - \Phi\left(R_{kl}^U(\vartheta)\right) + \Phi\left(R_{kl}^L(\vartheta)\right) \right\} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)^{\otimes 2}}{S(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{k,l} \left\{ 1 - \Phi\left(R_{kl}^U(\vartheta)\right) + \Phi\left(R_{kl}^L(\vartheta)\right) \right\} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)}{S(\vartheta)} \right\}^{\otimes 2} \right]$$

5.5 Simulation Studies

We conducted extensive simulation studies to assess the performance of the proposed methods. We considered both designs 1 and 2. Specifically, we generated a random sample of $N = 5{,}000$ individuals from the joint distribution of the trait value and genotype and identified the subset of all the individuals whose trait values are less than c_L or larger than c_U. We then selected a random sample of $n = 500$ individuals from that subset. By setting the genotypes of the unselected individuals to missing, we obtained the data under design 1; by deleting the unselected individuals altogether, we obtained the data under design 2. We evaluated both the full-likelihood and conditional-likelihood methods. These evaluations provided information about the relative efficiency of using full likelihood versus conditional likelihood under design 1 or equivalently the relative efficiency of design 1 versus design 2.

For comparison, we also evaluated the standard methods, which are based on the prospective likelihoods. For genotype-based analysis, the prospective likelihood is simply $\prod_{i=1}^{n} P(Y_i|G_i;\theta)$; (Wallace et al. 2006) for haplotype-based analysis, the prospective likelihood is the first term in (5.6) (Schaid et al. 2002).

In our first study, we generated the trait values from model (5.1) with $\alpha = 0$, $\sigma^2 = 1$ and $\beta = 0$, 0.1, 0.2, 0.3, 0.4 and 0.5. The potential causal variant was the minor allele. Under $\beta = 0$, the thresholds of -2.0, -1.5, -1.0, -0.5, 0.5 and 1.0 correspond approximately to the 2nd, 7th, 16th, 31st, 69th and 84th percentiles of the trait distribution, respectively. We considered three modes of inheritance: additive, dominant and recessive, and various values of the minor allele frequency (MAF). The genotypes were generated under HWE, and the analyses were performed both with and without this assumption. The results without the HWE assumption are summarized in Table 5.1. The results with HWE are similar and thus omitted.

Both the full and conditional likelihoods provide (virtually) unbiased estimators of the genetic effect and correct type I error. The standard error estimators accurately reflect the true variations, and the confidence intervals have proper coverages. The

conditional likelihood has nearly the same power as the full likelihood. As expected, the power is substantially higher under the additive and dominant models than under the recessive model (given the same MAF and the same effect size). The power increases as selection becomes more extreme. Also, the power tends to be higher when c_L and c_U are of the same distance from the population mean (as opposed to unequal distances). In practice, the population mean may be unknown or it may be easier to recruit subjects with high trait values than those with low trait values or vice versa. Thus, it may not be feasible to set c_L and c_U the same distance from the population mean.

In the presence of a causal variant, both the estimator of the genetic effect and the standard error estimator based on the prospective likelihood are biased upwards, and the coverages of the confidence intervals may be substantially below or above the desired levels. The prospective likelihood appears to preserve the type I error. The power of the prospective likelihood tends to be lower than that of the full and conditional likelihoods, especially when $(c_L, c_U) = (-2, 1)$ and under the recessive mode of inheritance. When $(c_L, c_U) = (-2, 1)$, the full and conditional likelihoods have power of approximately 75% to detect effect size of 0.3 under the additive and dominant models with MAF=0.05 and power of approximately 80% to detect effect size of 0.5 under the recessive model with MAF=0.2. By contrast, the prospective likelihood has less than 70% power in those two cases. The results in Table 5.1 pertain to likelihood-ratio tests. The results for Wallace et al.'s (2006) test, which is the score test based on the prospective likelihood, are virtually identical to those of the likelihood-ratio test (data not shown).

In the second study, we generated data in the same way as in the first study, but performed the analysis at a marker that is in linkage disequilibrium (LD) with the potential causal SNP. The results are shown in Table 5.2. The conclusions are essentially the same as in the first study. As expected, the power is decreased when testing is performed at a marker than at the candidate locus.

The third study was concerned with haplotype effects. We considered two SNPs

with varying degrees of LD. The 11 haplotype, i.e., the haplotype consisting of the minor allele at each site, had a potential effect on the trait value. Trait values were generated from model (5.5) with $\alpha = 0$, $\sigma^2 = 1$ and $\beta = 0$, 0.1, 0.2, 0.3, 0.4 and 0.5. We considered three modes of inheritance: additive, dominant and recessive. HWE was assumed in both the data generation and the analysis. Two types of analyses were performed: the first analysis compared the 11 haplotype to the other three haplotypes, and the second analysis compared haplotypes 11, 10, and 01 to haplotype 00. Some of the testing results are displayed in Figures 5.1 and 5.2.

The full and conditional likelihoods provide (virtually) unbiased estimators of haplotype effects. The standard error estimators are very accurate and the confidence intervals have correct coverages. The two methods have proper control of the type I error and very similar power. Not surprisingly, the power increases as LD becomes higher and as selection becomes more extreme. The prospective likelihood yields biased estimation of haplotype effects and inappropriate confidence intervals. As shown in Figure 5.1, the prospective likelihood is less powerful than the full and conditional likelihoods, especially under recessive mode of inheritance. Furthermore, the prospective likelihood yields inflated type I error for testing null haplotypes. The inflation of the type I error becomes more severe as the effect of the causal haplotype increases, as illustrated in Figure 5.2.

5.6 Discussion

The two designs considered in this chapter are quite general and flexible. Since the simulation studies indicated that conditional likelihoods are nearly as efficient as full likelihoods, one may simply adopt design 2 and retain the trait values for the genotyped individuals only. The choices of the selection thresholds do not require precise knowledge of the trait distribution, although the efficiency of the design will depend on which percentiles the thresholds correspond to. The likelihoods presented here can be easily modified to include a random sample, as in the original Slatkin's design, or

to allow several selection regions with different sampling probabilities. Although we have focused on normally distributed traits, our methods can be applied to any trait distributions.

Wallace et al. (2006) state that their test, which is the score test for a normal trait based on the prospective likelihood, is asymptotically equivalent to the score test based on the retrospective likelihood $\prod_{i=1}^{n} P(G_i|Y_i)$ under the null hypothesis of no causal variant. The proof given in their Appendix A requires that the average trait value of the selected individuals is an unbiased estimator of the population mean α. This assumption holds only when the upper and lower thresholds are of the same distance from the population mean and the low-trait value and high-trait value samples are of the same size. Even under such conditions, β and σ^2 cannot be estimated from the retrospective likelihood.

We have focused on the analysis of a single marker or a small set of markers. Association studies typically involve many markers, so a large number of tests may be performed. Adjustments for multiple testing can be made by permutation or Monte Carlo methods (Lin 2005).

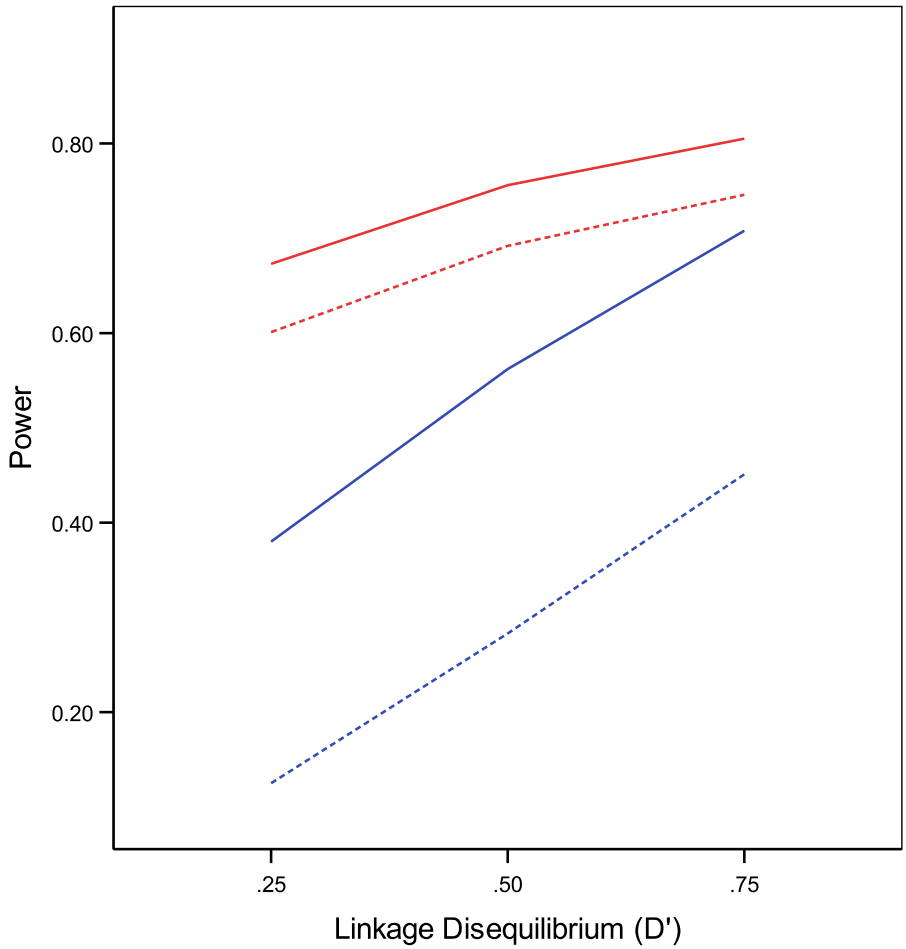


Figure 5.1: Empirical power for 2-SNP models as a function of linkage disequilibrium (D') between SNPs. The red curves correspond to a dominant model with effect size $\beta = .2$, and the blue curves correspond to a recessive model with $\beta = .3$. Solid curves pertain to the conditional analysis, while dotted curves pertain to the prospective analysis. The MAFs for the two SNPs are .3 and .4; $c_L = -2$ and $c_U = 1$. The nominal significance level is .05.

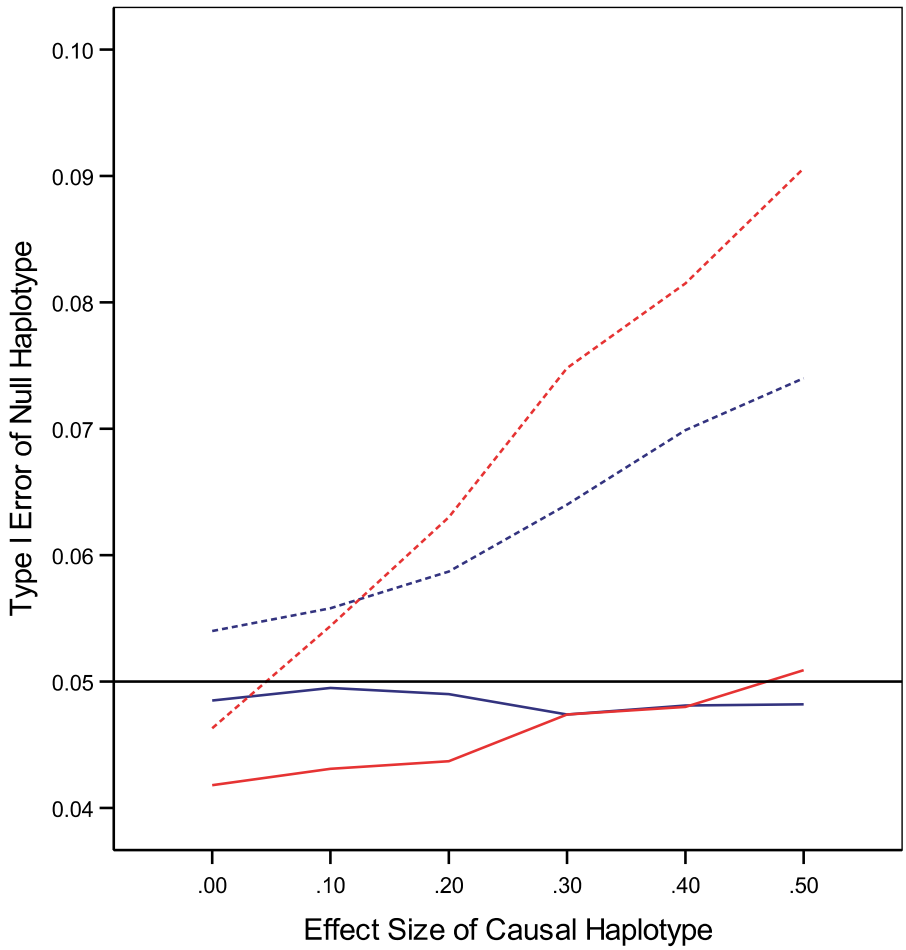


Figure 5.2: Empirical type I error for null haplotype 10 in a 2-SNP additive model as a function of effect size. The MAFs for the two SNPs are .3 and .4; the linkage disequilibrium (D') between the SNPs is 0.75. The red curves correspond to the case where $c_L = -2$ and $c_U = 1$, while the blue curves correspond to $c_L = -1$ and $c_U = 1$. Solid curves pertain to the conditional analysis, and dotted curves pertain to the prospective analysis. A solid black reference line is drawn at the nominal significance level of 0.05.

Table 5.1: Bias, standard error (SE), standard error estimate (SEE), coverage probability of the 95% confidence interval (CP) and power at the 0.05 nominal significance level. Each entry is based on 10,000 simulated datasets. c_L and c_U indicate the selection cutoffs for the sample of 500 genotyped subjects, out of a population of size 5,000. (a) Simulations from 1-SNP additive model with MAF of 0.05. (b) Simulations from 1-SNP dominant model with MAF of 0.05. (c) Simulations from 1-SNP recessive model with MAF of 0.2. Hardy-Weinberg equilibrium is not assumed.

(a)

			Full Likelihood					Conditional Likelihood					Prospective Likelihood				
β	c_L	c_U	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
0	-0.5	0.5	0.001	0.12	0.12	95.3	5.0	0.001	0.12	0.12	95.2	5.0	0.002	0.18	0.18	95.0	4.9
	-1.0	1.0	0.001	0.09	0.09	95.3	5.0	0.001	0.09	0.09	95.3	5.0	0.003	0.23	0.23	95.0	5.0
	-1.5	0.5	0.009	0.12	0.12	95.4	5.1	0.010	0.12	0.12	95.2	5.1	0.001	0.19	0.19	95.0	4.9
	-2.0	1.0	0.014	0.11	0.11	95.8	5.3	0.015	0.12	0.11	95.6	5.3	0.002	0.20	0.20	95.2	4.8
.20	-0.5	0.5	0.001	0.12	0.12	95.0	40.9	0.003	0.12	0.12	95.0	40.9	0.111	0.18	0.18	91.1	40.6
	-1.0	1.0	0.003	0.10	0.10	95.0	59.0	0.004	0.10	0.10	95.3	59.0	0.291	0.22	0.23	75.6	58.6

Continued on Next Page...

Table 5.1 (a) – Continued

			Full Likelihood					Conditional Likelihood					Prospective Likelihood				
β	c_L	c_U	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
	-1.5	0.5	0.011	0.13	0.13	95.0	40.0	0.014	0.13	0.13	94.8	40.0	0.079	0.15	0.17	95.8	35.8
	-2.0	1.0	0.016	0.13	0.13	95.0	42.2	0.020	0.13	0.13	94.9	42.2	0.084	0.14	0.18	97.1	34.5
.30	-0.5	0.5	0.002	0.12	0.12	95.2	72.9	0.004	0.12	0.12	95.5	73.0	0.159	0.17	0.18	86.1	72.8
	-1.0	1.0	0.003	0.10	0.10	95.4	90.3	0.004	0.10	0.10	95.3	90.2	0.403	0.20	0.22	55.7	90.0
	-1.5	0.5	0.010	0.13	0.13	94.6	70.7	0.014	0.14	0.13	94.7	70.6	0.084	0.14	0.17	96.2	66.5
	-2.0	1.0	0.016	0.14	0.14	94.7	75.2	0.022	0.14	0.14	95.0	75.1	0.076	0.12	0.17	98.5	68.6
.40	-0.5	0.5	0.003	0.12	0.12	94.8	92.8	0.005	0.12	0.12	95.2	92.8	0.199	0.17	0.18	80.9	92.7
	-1.0	1.0	0.007	0.10	0.10	95.1	99.0	0.009	0.11	0.10	95.0	98.2	0.500	0.19	0.22	33.5	99.0
	-1.5	0.5	0.008	0.14	0.13	94.3	91.4	0.013	0.14	0.14	94.6	91.4	0.074	0.13	0.16	97.6	89.5
	-2.0	1.0	0.015	0.14	0.14	94.2	93.3	0.021	0.15	0.14	94.3	93.4	0.048	0.11	0.16	99.3	90.4

(b)

			Full Likelihood					Conditional Likelihood					Prospective Likelihood				
β	c_L	c_U	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
0	-0.5	0.5	0.001	0.12	0.12	95.3	5.0	0.002	0.12	0.12	95.2	4.9	0.002	0.19	0.19	95.1	4.9
	-1.0	1.0	0.001	0.10	0.10	95.3	5.0	0.001	0.10	0.10	95.3	5.0	0.003	0.24	0.24	95.1	4.9
	-1.5	0.5	0.010	0.12	0.12	95.3	5.2	0.010	0.12	0.12	95.1	5.2	0.001	0.20	0.19	95.0	5.0
	-2.0	1.0	0.014	0.12	0.11	95.8	5.2	0.015	0.12	0.12	95.6	5.2	0.002	0.21	0.21	95.2	4.8
.20	-0.5	0.5	0.001	0.12	0.12	94.9	38.4	0.003	0.12	0.12	94.9	38.5	0.112	0.19	0.19	90.9	38.3
	-1.0	1.0	0.002	0.10	0.10	95.3	55.6	0.003	0.10	0.10	95.3	55.6	0.292	0.23	0.24	76.9	55.1
	-1.5	0.5	0.009	0.13	0.13	95.2	36.8	0.012	0.13	0.13	95.0	36.8	0.080	0.16	0.18	95.6	33.2
	-2.0	1.0	0.016	0.14	0.13	94.9	40.1	0.021	0.14	0.13	95.0	40.0	0.090	0.15	0.18	96.9	32.9
.30	-0.5	0.5	0.002	0.12	0.12	94.7	69.9	0.004	0.12	0.12	94.9	69.8	0.162	0.18	0.19	86.3	69.7
	-1.0	1.0	0.004	0.10	0.10	95.2	88.2	0.006	0.10	0.10	95.3	88.2	0.417	0.22	0.23	56.3	88.0
	-1.5	0.5	0.009	0.14	0.14	94.7	67.6	0.013	0.14	0.14	94.7	67.5	0.091	0.15	0.18	95.7	63.4
	-2.0	1.0	0.018	0.14	0.14	94.7	72.0	0.024	0.15	0.14	95.1	72.0	0.090	0.13	0.17	98.1	65.5
.40	-0.5	0.5	0.006	0.12	0.12	94.9	91.3	0.008	0.13	0.13	95.0	91.3	0.209	0.18	0.19	80.2	91.2
	-1.0	1.0	0.006	0.10	0.10	95.2	98.9	0.007	0.11	0.11	95.3	98.9	0.519	0.20	0.23	35.2	98.8
	-1.5	0.5	0.009	0.14	0.14	94.5	89.4	0.014	0.14	0.14	94.9	89.3	0.086	0.13	0.17	96.9	87.2
	-2.0	1.0	0.016	0.15	0.15	94.3	91.3	0.023	0.15	0.15	94.5	91.3	0.065	0.12	0.17	99.1	88.2

(c)

			Full Likelihood					Conditional Likelihood					Prospective Likelihood				
β	c_L	c_U	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
0	-0.5	0.5	-0.001	0.19	0.19	95.3	5.4	-0.001	0.19	0.19	95.3	5.4	-0.002	0.29	0.29	94.7	5.2
	-1.0	1.0	0.005	0.15	0.15	95.9	4.9	0.005	0.15	0.15	95.9	4.9	0.011	0.37	0.37	95.2	4.7
	-1.5	0.5	0.024	0.20	0.19	95.4	5.5	0.026	0.20	0.19	95.4	5.5	-0.000	0.30	0.30	94.9	5.1
	-2.0	1.0	0.041	0.20	0.19	96.0	5.4	0.043	0.20	0.19	95.8	5.5	0.004	0.32	0.32	95.5	4.5
.30	-0.5	0.5	0.005	0.19	0.19	95.1	37.6	0.008	0.20	0.19	95.4	37.7	0.160	0.28	0.29	91.9	37.1
	-1.0	1.0	0.011	0.16	0.16	96.0	54.6	0.014	0.16	0.16	95.8	54.6	0.417	0.33	0.36	79.0	53.7
	-1.5	0.5	0.028	0.21	0.21	95.0	36.6	0.035	0.22	0.21	95.0	36.5	0.095	0.22	0.27	98.0	28.9
	-2.0	1.0	0.041	0.22	0.22	94.6	39.0	0.050	0.23	0.22	94.7	38.9	0.089	0.19	0.27	99.4	25.1
.40	-0.5	0.5	0.005	0.20	0.19	94.5	58.1	0.009	0.20	0.19	95.0	58.0	0.201	0.27	0.28	90.2	57.3
	-1.0	1.0	0.018	0.17	0.16	95.5	79.0	0.022	0.17	0.17	95.6	79.0	0.524	0.31	0.35	67.9	78.2
	-1.5	0.5	0.024	0.22	0.22	93.9	56.6	0.031	0.22	0.22	94.4	56.5	0.087	0.20	0.26	98.8	48.2
	-2.0	1.0	0.037	0.23	0.23	94.0	60.2	0.048	0.23	0.23	94.2	60.1	0.066	0.17	0.25	99.7	46.0
.50	-0.5	0.5	0.010	0.20	0.19	94.6	77.7	0.014	0.20	0.20	95.2	77.7	0.242	0.26	0.28	88.3	77.1
	-1.0	1.0	0.021	0.17	0.17	95.6	93.5	0.026	0.18	0.17	95.6	93.4	0.600	0.28	0.34	57.1	93.1
	-1.5	0.5	0.018	0.22	0.22	94.2	74.9	0.027	0.22	0.22	94.5	74.7	0.068	0.18	0.25	99.2	68.0
	-2.0	1.0	0.027	0.22	0.23	94.1	79.0	0.039	0.23	0.23	94.4	78.9	0.027	0.15	0.24	99.8	68.0

Table 5.2: Type I error and power of marker SNP in LD (D'=0.9) with causal SNP in a 2-SNP model. The additive and dominant models have causal and marker SNPs with MAFs of .05 and .06; the recessive model has SNPs with MAFs of .2 and .25. Hardy-Weinberg equilibrium is not assumed.

		Additive				Dominant				Recessive			
c_L	c_U	β	Full	Cond	Pros	β	Full	Cond	Pros	β	Full	Cond	Pros
-0.5	0.5	0.0	5.3	5.2	5.2	0.0	5.2	5.3	5.3	0.0	5.1	5.1	5.0
-1.0	1.0		5.2	5.2	5.1		5.2	5.2	5.2		5.7	5.7	5.6
-1.5	0.5		4.7	4.6	4.7		4.5	4.5	4.7		5.4	5.4	5.0
-2.0	1.0		5.5	5.5	5.5		5.3	5.4	5.5		5.1	5.2	4.5
-0.5	0.5	0.2	29.4	29.5	29.3	0.2	28.0	28.1	27.9	0.3	20.2	20.2	20.0
-1.0	1.0		42.7	42.7	42.3		40.1	40.1	39.9		29.1	29.1	28.8
-1.5	0.5		28.2	28.1	24.9		26.0	26.0	23.1		19.5	19.4	15.4
-2.0	1.0		29.5	29.6	24.0		28.1	28.1	23.3		19.8	19.7	13.6
-0.5	0.5	0.3	55.3	55.3	55.1	0.3	51.9	51.8	51.6	0.4	30.2	30.3	30.0
-1.0	1.0		76.0	76.0	75.8		72.7	72.7	72.3		45.0	45.0	44.4
-1.5	0.5		54.0	54.0	49.9		49.9	50.0	46.3		29.5	29.4	24.5
-2.0	1.0		56.2	56.2	49.5		52.4	52.3	46.1		31.7	31.5	23.5

-0.5	0.5	0.4	79.4	79.4	79.1	0.4	75.6	75.7	75.5	0.5	44.1	44.1	43.8
-1.0	1.0		93.7	93.7	93.5		92.0	92.0	91.8		63.3	63.3	62.6
-1.5	0.5		75.8	75.7	72.5		72.6	72.7	69.7		42.0	41.8	36.5
-2.0	1.0		78.8	78.7	73.8		75.8	75.7	71.1		45.1	45.0	35.8

Note: Each entry is based on 10000 simulated datasets. c_L and c_U indicate the selection cutoffs for the sample of 500 genotyped subjects, out of a population of size 5,000. Full, Cond., and Pros. stand for full, conditional, and prospective likelihood analysis respectively.

Chapter 6

Association Mapping of QTLs with General Two-Phase Designs

In the previous chapter, we discussed the use of selective genotyping designs to reduce cost and improve efficiency for genetic association studies. We mentioned methods for various selective genotyping designs (Slatkin 1999; Chen et al. 2005; Wallace et al. 2006) and noted the general loss of information for many of the approaches. The maximum-likelihood methods we proposed were shown to have numerous advantages over the alternatives, including negligible bias, increased power, and appropriate coverage of confidence intervals.

As selective genotyping designs are a subclass of ODS designs, we can generalize our approach to some commonly discussed ODS designs. Two-phase sampling has been discussed in Lawless et al. (1999) and Scott and Wild (2000), who demonstrated that computation of maximum likelihood estimators under semiparametric regression models was feasible for a wide range of designs. Breslow et al. (2003) extend these results by providing asymptotic theory for these estimators. Lawless et al. (1999) consider semiparametric methods which treat the marginal distribution of covariates nonparametrically. These are directly applicable to testing for genotype-disease association; however, this is not true when testing haplotypes, which are not directly measured.

A further important extension is the inclusion of environmental factors in our models. This is vital in association studies of complex quantitative traits, which are influenced by a combination of many environmental and genetic factors. Incorporating covariates into appropriate likelihoods for selective genotyping designs involves a higher degree of complexity, however, due to the presence of the covariate distribution. None of the previous selective genotyping methods are formulated to be applicable to studies which include covariates.

In this chapter, we extend the approach of the previous chapter to more general selective-genotyping designs, including those which allow for environmental covariates. Designs 1 and 2 from the previous chapter selected samples from the two tails of the distribution with equal selection probabilities. We generalize this to the two-phase design of Lawless et al. (1999), which allows for more strata, and arbitrary selection probabilities. We modify it for haplotype-disease association mapping, and then further extend it to test for both haplotype and covariate associations.

The likelihoods derived here make full use of the available data and properly reflect the selection process. The corresponding inference procedures are valid and efficient. For designs including covariates, we profile the observed-data likelihoods over the (possibly infinite-dimensional) nuisance parameter of the covariate distribution. When covariates and genotypes are independent, the resulting profile likelihoods are shown to satisfy conditions similar to those presented in Breslow et al. (2003) and hence the maximum likelihood estimators are asymptotically normal and efficient. The properties of the proposed estimators and their advantages over other approaches are demonstrated through simulation studies.

6.1 Two-Phase Selective Genotyping Design

As in Chapter 5, we are interested in characterizing the relationship between a response Y and haplotype (and covariates) by using maximum likelihood methods. We consider a more general selective genotyping design than previously by allowing

for multiple strata. This two-phase sampling design is described in Lawless et al. (1999). In the previous chapter we discuss the characterization of both genotype and haplotype association. The methods discussed in Lawless et al. (1999) cover the case of genotype association, so we focus on haplotype association here. We begin by examining the case with no covariates to give the basic setup; the extension to covariates is described later.

We sample N observations from the population. Let Y_i denote the trait value of the ith individual, and G_i is the corresponding genotype denoting the number of minor alleles at each locus. The range of Y is partitioned into J strata $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_J$. We assume that not all subjects are fully observed, and define $R_i = 1$ if (Y_i, G_i) is fully observed, and 0 if some information on (Y_i, G_i) is missing. If $R_i = 0$ then the only information retained is the identity of the stratum that Y_i is in. The number of subjects in stratum $\mathcal{C}_j$ is written as N_j, and the number of fully observed subjects is n_j, with $N = \sum_j N_j$ and $n = \sum_j n_j$.

Observations are selected for genotyping, and hence fully observed, according to a variable probability sampling scheme. Subjects are inspected sequentially, and their stratum is identified. When the trait value Y_i of an individual falls into $\mathcal{C}_j$, that individual is selected for genotyping ($R_i = 1$) with specified probability p_j. There are two possibilities for stopping rules under this sampling scheme, namely, to either fix N or n beforehand, and inspect units until this cutoff is reached. We focus on the latter case, as it is more applicable to genetic association studies, where the budget may be allocated to genotype a prespecified number of subjects. In this situation the number of subjects genotyped in each stratum, n_j, is random.

While the genotype and trait values constitute the observed data for individuals who are fully observed, we are interested in testing for association between the diplotype H and Y of an individual. This is characterized by the conditional density function $P(Y_i|H_i;\theta)$ indexed by a set of parameters θ. $P(Y_i|H_i;\theta)$ may take the form of the linear regression model

$$Y_i = \alpha + \beta Z(H_i) + \epsilon_i, \tag{6.1}$$

where ϵ_i is zero-mean normal with variance σ^2. The function $Z(H_i)$ is a haplotype score and will depend on the mode of inheritance; for a specific haplotype in an additive model, it would be the number of occurrences of the haplotype for each individual. In this model, $\theta = (\alpha, \beta, \sigma^2)$. Let $S(G)$ denote the set of diplotypes compatible with a given genotype G.

Because haplotypes are not directly observed, it is necessary to impose some restrictions, such as Hardy-Weinberg Equilibrium (HWE), on the diplotype distribution. For $k = 1, \ldots, K$, let h_k denote the kth possible haplotype in the population and let π_k denote the population frequency of h_k. Under HWE, $P[H_i = (h_k, h_l)] = \pi_k \pi_l$ $(k, l = 1, \ldots, K)$. We denote the diplotype probability function by $P(H_i; \pi)$, where $\pi = (\pi_1, \ldots, \pi_K)$.

Under this design, the data consist of (Y_i, G_i) $(i = 1, \ldots, n)$ and N_j $(j = 1, \ldots, J)$. We consider a likelihood corresponding to the full semiparametric likelihood described in Lawless et al. (1999). This is defined as $L_F(\theta, \pi)$ and given by:

$$\prod_{j=1}^{J} \left[\prod_{i=1}^{n_j} \sum_{H \in S(G_i)} P(Y_i|H;\theta) P(H;\pi) \right] \left[\sum_{G} \sum_{H \in S(G)} P(Y \in \mathcal{C}_j|H;\theta) P(H;\pi) \right]^{N_j - n_j} \tag{6.2}$$

Alternately, if stratum values are not retained for the subjects who are not genotyped, we can write down the conditional likelihood

$$L_C(\theta, \pi) = \prod_{i:R_i=1} P(Y_i, G_i|R_i = 1) \tag{6.3}$$

where R_i is the indicator for whether an individual is genotyped or not. If we let $p_j = P(R_i = 1|Y_i \in \mathcal{C}_j)$ and $\delta_{ij} = I(Y_i \in \mathcal{C}_j)$, then this conditional likelihood can further be written as

$$\prod_{i=1}^{n} \frac{\sum_{j=1}^{J} \delta_{ij} p_j \sum_{H \in S(G_i)} P(Y_i|H;\theta) P(H;\pi)}{\sum_{j=1}^{J} \sum_{G} \sum_{H \in S(G)} p_j P(Y_i \in \mathcal{C}_j|H;\theta) P(H;\pi)} \tag{6.4}$$

These likelihoods can be maximized using Newton-Raphson algorithms, for which details are given in the next two sections.

6.2 Newton-Raphson Algorithm to Maximize (6.2)

Under the linear regression model with strata defined as (L_j, U_j), the log-likelihood corresponding to (6.2) is, up to a constant,

$$\sum_{j=1}^{J}\left[\sum_{i=1}^{n_j}\log\sum_{(h_k,h_l)\in S(G_i)}(2\pi\sigma^2)^{-1/2}\exp\left\{-(Y_i-\beta^T Z(h_k,h_l))^2/(2\sigma^2)\right\}\pi_k\pi_l\right.$$
$$\left.+(N_j-n_j)\log\sum_{k,l}\left[\Phi\left(\frac{U_j-\beta^T Z(h_k,h_l)}{\sigma}\right)-\Phi\left(\frac{L_j-\beta^T Z(h_k,h_l)}{\sigma}\right)\right]\pi_k\pi_l\right].$$

For notational convenience, denote σ^2 as v. Let

$$\begin{aligned}
Q_{ikl}(\theta) &= \exp\left\{-(Y_i-\beta^T Z(h_k,h_l))^2/(2v)\right\},\\
R^L_{klj}(\theta) &= \{L_j-\beta^T Z(h_k,h_l)\}/\sqrt{v},\\
R^U_{klj}(\theta) &= \{U_j-\beta^T Z(h_k,h_l)\}/\sqrt{v},\\
S_j(\theta,\pi) &= \sum_{k,l}\left\{\Phi\left(R^U_{klj}(\theta)\right)-\Phi\left(R^L_{klj}(\theta)\right)\right\}\pi_k\pi_l.
\end{aligned}$$

Also, let $a^{\otimes 2}=aa^T$ and let ϕ be the standard normal density function. Then

$$\frac{\partial\ell_F(\theta,\pi)}{\partial v}=-\frac{n}{2v}+\sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)\pi_k\pi_l\frac{(Y_i-\beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)\pi_k\pi_l}$$
$$-\sum_{j=1}^{J}(N_j-n_j)\frac{\sum_{k,l}\left\{\phi\left(R^U_{klj}(\theta)\right)R^U_{klj}(\theta)-\phi\left(R^L_{klj}(\theta)\right)R^L_{klj}(\theta)\right\}\frac{\pi_k\pi_l}{2v}}{S_j(\theta,\pi)}$$

$$\frac{\partial\ell_F(\theta,\pi)}{\partial\beta}=\sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)}\frac{Q_{ikl}(\theta)}{v}(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)\pi_k\pi_l}$$
$$-\sum_{j=1}^{J}(N_j-n_j)\frac{\sum_{k,l}\left\{\phi\left(R^U_{klj}(\theta)\right)-\phi\left(R^L_{klj}(\theta)\right)\right\}\frac{Z(h_k,h_l)}{\sqrt{v}}\pi_k\pi_l}{S_j(\theta,\pi)}$$

$$\frac{\partial\ell_F(\theta,\pi)}{\partial\pi}=\sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)D^{kl}_{\pi}}{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)\pi_k\pi_l}$$
$$+\sum_{j=1}^{J}(N_j-n_j)\frac{\sum_{k,l}\left\{\Phi\left(R^U_{klj}(\theta)\right)-\Phi\left(R^L_{klj}(\theta)\right)\right\}D^{kl}_{\pi}}{S_j(\theta,\pi)}$$

Let Δ_{ij} be the Kronecker delta which is 1 if $i = j$ and 0 otherwise. Here we define D_π^{kl} as the vector of derivatives of $\pi_k\pi_l$ with respect to π which has as its mth element $\pi_k(\Delta_{lm} - \Delta_{lK}) + \pi_l(\Delta_{km} - \Delta_{kK})$. This takes into account the constraint that $\sum_k \pi_k = 1$. Also define $M_{\pi\pi}^{kl}$ as the matrix of second derivatives of $\pi_k\pi_l$. This has as its mnth element $(\Delta_{lm} - \Delta_{lK})(\Delta_{kn} - \Delta_{kK}) + (\Delta_{ln} - \Delta_{lK})(\Delta_{km} - \Delta_{kK})$.

$$\frac{\partial^2 \ell_F(\theta,\pi)}{\partial v^2} = \frac{n}{2v^2} + \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^4}{4v^4} - \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{v^3} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right\}^2 \right]$$
$$- \sum_{j=1}^{J} (N_j - n_j) \left[\left\{ \frac{\sum_{k,l} \left(\phi(R_{klj}^U(\theta)) R_{klj}^U(\theta) - \phi(R_{klj}^L(\theta)) R_{klj}^L(\theta) \right) \frac{\pi_k\pi_l}{2v}}{S_j(\theta,\pi)} \right\}^2 \right.$$
$$+ \frac{\sum_{k,l} \left\{ \phi(R_{klj}^U(\theta)) \left(\left(R_{klj}^U(\theta)^3 \right) - 3R_{klj}^U(\theta) \right) \right\} \frac{\pi_k\pi_l}{4v^2}}{S_j(\theta,\pi)}$$
$$\left. - \frac{\sum_{k,l} \left\{ \phi(R_{klj}^L(\theta)) \left(\left(R_{klj}^L(\theta)^3 \right) - 3R_{klj}^L(\theta) \right) \right\} \frac{\pi_k\pi_l}{4v^2}}{S_j(\theta,\pi)} \right]$$

$$\frac{\partial^2 \ell_F(\theta,\pi)}{\partial v \partial \beta} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) Z(h_k,h_l) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^3}{2v^3} - \frac{(Y_i - \beta^T Z(h_k,h_l))}{v^2} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right.$$
$$- \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right\}$$
$$\left. \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right\} \right]$$
$$- \sum_{j=1}^{J} (N_j - n_j) \left[\left\{ \frac{\sum_{k,l} \left(\phi(R_{klj}^U(\theta)) R_{klj}^U(\theta) - \phi(R_{klj}^L(\theta)) R_{klj}^L(\theta) \right) \frac{\pi_k\pi_l}{2v}}{S_j(\theta,\pi)} \right\} \right.$$
$$\left\{ \frac{\sum_{k,l} \left(\phi(R_{klj}^U(\theta)) - \phi(R_{klj}^L(\theta)) \right) \pi_k\pi_l \frac{Z(h_k,h_l)}{\sqrt{v}}}{S_j(\theta,\pi)} \right\}$$
$$\left. + \frac{\sum_{k,l} \left\{ \phi(R_{klj}^U(\theta)) \left(R_{klj}^U(\theta)^2 - 1 \right) - \phi(R_{klj}^L(\theta)) \left(R_{klj}^L(\theta)^2 - 1 \right) \right\} \frac{\pi_k\pi_l}{2v^{3/2}} Z(h_k,h_l)}{S_j(\theta,\pi)} \right]$$

$$\frac{\partial^2 \ell_F(\theta,\pi)}{\partial(\beta\beta^T)} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) \left\{ \frac{(Y_i-\beta^T Z(h_k,h_l))^2}{v^2} - v^{-1} \right\} Z^{\otimes 2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) \frac{(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right\}^{\otimes 2} \right]$$
$$- \sum_{j=1}^{J} (N_j - n_j) \left[\left\{ \frac{\sum_{k,l} \left(\phi(R^U_{klj}(\theta)) - \phi(R^L_{klj}(\theta)) \right) \pi_k\pi_l \frac{Z(h_k,h_l)}{\sqrt{v}}}{S_j(\theta,\pi)} \right\}^{\otimes 2} \right.$$
$$\left. + \frac{\sum_{k,l} \left\{ \phi(R^U_{klj}(\theta)) R^U_{klj}(\theta) - \phi(R^L_{klj}(\theta)) R^L_{klj}(\theta) \right\} \pi_k\pi_l \left(\frac{Z(h_k,h_l)}{\sqrt{v}} \right)^{\otimes 2}}{S_j(\theta,\pi)} \right]$$

$$\frac{\partial^2 \ell_F(\theta,\pi)}{\partial v \partial \pi} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) D^{kl}_{\pi} \frac{(Y_i-\beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} \frac{(Y_i-\beta^T Z(h_k,h_l))^2}{2v^2} Q_{ikl}(\theta)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right\} \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta) D^{kl}_{\pi}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} \right\} \right]$$
$$+ \sum_{j=1}^{J} (N_j - n_j) \left[\left\{ \frac{\sum_{k,l} \left\{ \phi\left(R^U_{klj}(\theta)\right) R^U_{klj}(\theta) - \phi\left(R^L_{klj}(\theta)\right) R^L_{klj}(\theta) \right\} \frac{\pi_k\pi_l}{2v}}{S_j(\theta,\pi)} \right\} \right.$$
$$\left\{ \frac{\sum_{k,l} \left\{ \Phi\left(R^U_{klj}(\theta)\right) - \Phi\left(R^L_{klj}(\theta)\right) \right\} D^{kl}_{\pi}}{S_j(\theta,\pi)} \right\}$$
$$\left. - \frac{\sum_{k,l} \left\{ \phi\left(R^U_{klj}(\theta)\right) R^U_{klj}(\theta) - \phi\left(R^L_{klj}(\theta)\right) R^L_{klj}(\theta) \right\} \frac{D^{kl}_{\pi}}{2v}}{S_j(\theta,\pi)} \right]$$

$$\frac{\partial^2 \ell_F(\theta,\pi)}{\partial\beta\partial\pi^T} = \sum_{i=1}^{n}\left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\frac{(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{v}(D_\pi^{kl})^T}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l}\right.$$
$$\left. - \left\{\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\frac{(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l}\right\}\left\{\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)D_\pi^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l}\right\}^T\right]$$
$$+\sum_{j=1}^{J}(N_j-n_j)\left[\left\{\frac{\sum_{k,l}\left\{\phi\left(R_{klj}^U(\theta)\right)-\phi\left(R_{klj}^L(\theta)\right)\right\}\frac{Z(h_k,h_l)}{\sqrt{v}}\pi_k\pi_l}{S_j(\theta,\pi)}\right\}\right.$$
$$\left\{\frac{\sum_{k,l}\left\{\Phi\left(R_{klj}^U(\theta)\right)-\Phi\left(R_{klj}^L(\theta)\right)\right\}D_\pi^{kl}}{S_j(\theta,\pi)}\right\}^T$$
$$\left. - \frac{\sum_{k,l}\left\{\phi\left(R_{klj}^U(\theta)\right)-\phi\left(R_{klj}^L(\theta)\right)\right\}\frac{Z(h_k,h_l)}{\sqrt{v}}D_\pi^{kl}}{S_j(\theta,\pi)}\right]$$

$$\frac{\partial^2 \ell_F(\theta,\pi)}{\partial(\pi\pi^T)} = \sum_{i=1}^{n}\left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)M_{\pi\pi}^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l} - \left\{\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)D_\pi^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\theta)\pi_k\pi_l}\right\}^{\otimes 2}\right]$$
$$+\sum_{j=1}^{J}(N_j-n_j)\left[\frac{\sum_{k,l}\left\{\Phi\left(R_{klj}^U(\theta)\right)-\Phi\left(R_{klj}^L(\theta)\right)\right\}M_{\pi\pi}^{kl}}{S_j(\theta,\pi)}\right.$$
$$\left. - \left\{\frac{\sum_{k,l}\left\{\Phi\left(R_{klj}^U(\theta)\right)-\Phi\left(R_{klj}^L(\theta)\right)\right\}D_\pi^{kl}}{S_j(\theta,\pi)}\right\}^{\otimes 2}\right]$$

6.3 Newton-Raphson Algorithm to Maximize (6.4)

Under the linear regression model with strata of the form (L_j, U_j), (6.4) becomes

$$\prod_{i=1}^{n}\frac{\sum_{j=1}^{J}\delta_{ij}p_j\sum_{(h_k,h_l)\in S(G_i)}(2\pi\sigma^2)^{-1/2}\exp\left\{-(Y_i-\beta^T Z(h_k,h_l))^2/(2\sigma^2)\right\}\pi_k\pi_l}{\sum_{j=1}^{J}p_j\sum_{k,l}\left[\Phi\left(\frac{U_j-\beta^T Z(h_k,h_l)}{\sigma}\right)-\Phi\left(\frac{L_j-\beta^T Z(h_k,h_l)}{\sigma}\right)\right]\pi_k\pi_l}.$$

To incorporate the constraints that $\sum_{k=1}^{K}\pi_k = 1$ and $\pi_k > 0$ $(k = 1,\ldots,K)$ into the calculations, we define $\pi_k^* = \pi_k/\pi_K$ and $\eta_k = \log\pi_k^*$. For notational convenience, denote σ^2 as v. Let $\eta = (\eta_1,\ldots,\eta_{K-1})$ and $\vartheta = (\beta, v, \eta)$. Then the log-likelihood is,

up to a constant,

$$\ell_C(\vartheta) = -\frac{n}{2}\log v + \sum_{i=1}^{n}\log \sum_{(h_k,h_l)\in S(G_i)} \exp\left\{-\frac{(Y_i-\beta^T Z(h_k,h_l))^2}{2v} + \eta^T W(h_k,h_l)\right\}$$
$$-\sum_{i=1}^{n}\log\sum_{j=1}^{J} p_j \sum_{k,l} e^{\eta^T W(h_k,h_l)}\left\{\Phi\left(\frac{U_j-\beta^T Z(h_k,h_l)}{\sqrt{v}}\right)-\Phi\left(\frac{L_j-\beta^T Z(h_k,h_l)}{\sqrt{v}}\right)\right\},$$

where

$$W(h_k,h_l) = \begin{bmatrix} I(h_k=h_1)+I(h_l=h_1) \\ \vdots \\ I(h_k=h_{K-1})+I(h_l=h_{K-1}) \end{bmatrix}.$$

Let

$$\begin{aligned}
Q_{ikl}(\vartheta) &= \exp\left\{-(Y_i-\beta^T Z(h_k,h_l))^2/(2v) + \eta^T W(h_k,h_l)\right\}, \\
R^L_{klj}(\vartheta) &= \{L_j-\beta^T Z(h_k,h_l)\}/\sqrt{v}, \\
R^U_{klj}(\vartheta) &= \{U_j-\beta^T Z(h_k,h_l)\}/\sqrt{v}, \\
S(\vartheta) &= \sum_{j=1}^{J} p_j \sum_{k,l}\left\{\Phi\left(R^U_{klj}(\vartheta)\right)-\Phi\left(R^L_{klj}(\vartheta)\right)\right\} e^{\eta^T W(h_k,h_l)}.
\end{aligned}$$

Also, let $a^{\otimes 2} = aa^T$ and let ϕ be the standard normal density function. Then

$$\frac{\partial \ell_C(\vartheta)}{\partial v} = -\frac{n}{2v} + \sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\frac{(Y_i-\beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}$$
$$+\sum_{i=1}^{n}\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\phi\left(R^U_{klj}(\vartheta)\right)R^U_{klj}(\vartheta)-\phi\left(R^L_{klj}(\vartheta)\right)R^L_{klj}(\vartheta)\right\}\frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)}$$

$$\frac{\partial \ell_C(\vartheta)}{\partial \beta} = \sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)}\frac{Q_{ikl}(\vartheta)}{v}(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}$$
$$+\sum_{i=1}^{n}\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\phi\left(R^U_{klj}(\vartheta)\right)-\phi\left(R^L_{klj}(\vartheta)\right)\right\}\frac{Z(h_k,h_l)}{\sqrt{v}}e^{\eta^T W(h_k,h_l)}}{S(\vartheta)}$$

$$\frac{\partial \ell_C(\vartheta)}{\partial \eta} = \sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}$$
$$-\sum_{i=1}^{n}\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\Phi\left(R^U_{klj}(\vartheta)\right)-\Phi\left(R^L_{klj}(\vartheta)\right)\right\} e^{\eta^T W(h_k,h_l)}W(h_k,h_l)}{S(\vartheta)}$$

$$\frac{\partial^2 \ell_C(\vartheta)}{\partial v^2} = \frac{n}{2v^2} + \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^4}{4v^4} - \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{v^3} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}^2 \right]$$
$$+ \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi(R^U_{klj}(\vartheta)) \left(R^U_{klj}(\vartheta)^3 - 3R^U_{klj}(\vartheta) \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{4v^2}}{S(\vartheta)} \right.$$
$$- \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi(R^L_{klj}(\vartheta)) \left(R^L_{klj}(\vartheta)^3 - 3R^L_{klj}(\vartheta) \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{4v^2}}{S(\vartheta)}$$
$$\left. + \left\{ \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left(\phi(R^U_{klj}(\vartheta)) R^U_{klj}(\vartheta) - \phi(R^L_{klj}(\vartheta)) R^L_{klj}(\vartheta) \right) \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)} \right\}^2 \right]$$

$$\frac{\partial^2 \ell_C(\vartheta)}{\partial v \partial \beta} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) Z(h_k,h_l) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^3}{2v^3} - \frac{(Y_i - \beta^T Z(h_k,h_l))}{v^2} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$- \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \right]$$
$$+ \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi(R^U_{klj}(\vartheta)) \left(R^U_{klj}(\vartheta)^2 - 1 \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v^{3/2}} Z(h_k,h_l)}{S(\vartheta)} \right.$$
$$- \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi(R^L_{klj}(\vartheta)) \left(R^L_{klj}(\vartheta)^2 - 1 \right) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v^{3/2}} Z(h_k,h_l)}{S(\vartheta)}$$
$$+ \left\{ \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left(\phi(R^U_{klj}(\vartheta)) R^U_{klj}(\vartheta) - \phi(R^L_{klj}(\vartheta)) R^L_{klj}(\vartheta) \right) \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left(\phi(R^U_{klj}(\vartheta)) - \phi(R^L_{klj}(\vartheta)) \right) e^{\eta^T W(h_k,h_l)} \frac{Z(h_k,h_l)}{\sqrt{v}}}{S(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell_C(\vartheta)}{\partial(\beta\beta^T)} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \left\{ \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{v^2} - v^{-1} \right\} Z^{\otimes 2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \beta^T Z(h_k,h_l)) Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\}^{\otimes 2} \right]$$
$$+ \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi(R^U_{klj}(\vartheta)) R^U_{klj}(\vartheta) - \phi(R^L_{klj}(\vartheta)) R^L_{klj}(\vartheta) \right\} e^{\eta^T W(h_k,h_l)} \left(\frac{Z(h_k,h_l)}{\sqrt{v}} \right)^{\otimes 2}}{S(\vartheta)} \right.$$
$$\left. + \left\{ \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left(\phi(R^U_{klj}(\vartheta)) - \phi(R^L_{klj}(\vartheta)) \right) e^{\eta^T W(h_k,h_l)} \frac{Z(h_k,h_l)}{\sqrt{v}}}{S(\vartheta)} \right\}^{\otimes 2} \right]$$

$$\frac{\partial^2 \ell_C(\vartheta)}{\partial v \partial \eta} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l) \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} \frac{(Y_i - \beta^T Z(h_k,h_l))^2}{2v^2} Q_{ikl}(\vartheta)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)} \right\} \right]$$
$$+ \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi\left(R^U_{klj}(\vartheta)\right) R^U_{klj}(\vartheta) - \phi\left(R^L_{klj}(\vartheta)\right) R^L_{klj}(\vartheta) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v} W(h_k,h_l)}{S(\vartheta)} \right.$$
$$\left. - \left\{ \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \phi\left(R^U_{klj}(\vartheta)\right) R^U_{klj}(\vartheta) - \phi\left(R^L_{klj}(\vartheta)\right) R^L_{klj}(\vartheta) \right\} \frac{e^{\eta^T W(h_k,h_l)}}{2v}}{S(\vartheta)} \right\} \right.$$
$$\left. \left\{ \frac{\sum_{j=1}^{J} p_j \sum_{k,l} \left\{ \Phi\left(R^U_{klj}(\vartheta)\right) - \Phi\left(R^L_{klj}(\vartheta)\right) \right\} e^{\eta^T W(h_k,h_l)} W(h_k,h_l)}{S(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell_C(\vartheta)}{\partial\beta\partial\eta^T} = \sum_{i=1}^{n}\left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\frac{(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{v}W(h_k,h_l)^T}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}\right.$$
$$\left. - \left\{\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\frac{(Y_i-\beta^T Z(h_k,h_l))Z(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}\right\}\left\{\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}\right\}^T\right]$$
$$+\sum_{i=1}^{n}\left[\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\phi\left(R^U_{klj}(\vartheta)\right)-\phi\left(R^L_{klj}(\vartheta)\right)\right\}\frac{Z(h_k,h_l)}{\sqrt{v}}e^{\eta^T W(h_k,h_l)}W(h_k,h_l)^T}{S(\vartheta)}\right.$$
$$- \left\{\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\phi\left(R^U_{klj}(\vartheta)\right)-\phi\left(R^L_{klj}(\vartheta)\right)\right\}\frac{Z(h_k,h_l)}{\sqrt{v}}e^{\eta^T W(h_k,h_l)}}{S(\vartheta)}\right\}$$
$$\left.\left\{\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\Phi\left(R^U_{klj}(\vartheta)\right)-\Phi\left(R^L_{klj}(\vartheta)\right)\right\}e^{\eta^T W(h_k,h_l)}W(h_k,h_l)}{S(\vartheta)}\right\}^T\right]$$

$$\frac{\partial^2 \ell_C(\vartheta)}{\partial(\eta\eta^T)} = \sum_{i=1}^{n}\left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)W(h_k,h_l)^{\otimes 2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}\right.$$
$$\left. - \left\{\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)W(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)}\right\}^{\otimes 2}\right]$$
$$-\sum_{i=1}^{n}\left[\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\Phi\left(R^U_{klj}(\vartheta)\right)-\Phi\left(R^L_{klj}(\vartheta)\right)\right\}e^{\eta^T W(h_k,h_l)}W(h_k,h_l)^{\otimes 2}}{S(\vartheta)}\right.$$
$$\left. - \left\{\frac{\sum_{j=1}^{J} p_j \sum_{k,l}\left\{\Phi\left(R^U_{klj}(\vartheta)\right)-\Phi\left(R^L_{klj}(\vartheta)\right)\right\}e^{\eta^T W(h_k,h_l)}W(h_k,h_l)}{S(\vartheta)}\right\}^{\otimes 2}\right]$$

6.4 Two-Phase Selective Genotyping with Environmental Factors

The likelihoods described above can be further generalized to incorporate covariate effects. Suppose we now have a set of covariates X_i which are observed only for the n genotyped individuals, along with response Y_i and genotype G_i. We can modify $P(Y_i|H_i, X_i; \theta)$ by including covariate effects in the regression model in (6.1)

$$Y_i = \alpha + \beta Z(H_i) + \gamma X_i + \epsilon_i, \tag{6.5}$$

so that θ now consists of $(\alpha, \beta, \gamma, \sigma^2)$.

The likelihoods given in (6.2) and (6.4) then become

$$L_F(\theta,\pi) = \prod_{j=1}^{J}\left[\prod_{i=1}^{n_j}\sum_{H\in S(G_i)} P(Y_i|X_i,H;\theta)P(H;\pi)P(X_i|G_i)\right] \left[\sum_{X,G}\sum_{H\in S(G)} P(Y_i\in\mathcal{C}_j|X,H;\theta)P(H;\pi)P(X|G)\right]^{N_j-n_j} \tag{6.6}$$

and

$$\prod_{i=1}^{n}\frac{\sum_{j=1}^{J}\delta_{ij}p_j\sum_{H\in S(G_i)}P(Y_i|H;\theta)P(H;\pi)P(X_i|G_i)}{\sum_{j=1}^{J}\sum_{X,G}\sum_{H\in S(G)}p_jP(Y_i\in\mathcal{C}_j|H;\theta)P(H;\pi)P(X|G)} \tag{6.7}$$

respectively.

The covariate distribution may be an infinite-dimensional nuisance parameter in the case of a continuous covariate. In this case, we construct the profile likelihood for θ, maximizing over $P(X|G)$ using the methods of Scott and Wild (1997). Let $x_{g,1},\ldots,x_{g,n_g}$ be distinct observed covariate values with $G=g$. Let ζ_{gk} be the probability of $X=x_{gk}$ given that $G=g$ for $k=1,\ldots,n_g$. Write n_{gk} as the number of individuals with $X=x_{gk}$ and $G=g$, and let n_{g+} be the total number of individuals with a given genotype. Finally, define

$$\eta_j(x,g;\theta,\pi) = \sum_{H\in S(g)} P(Y\in\mathcal{C}_j|x,H;\theta)P(H;\pi).$$

Then the log-likelihood corresponding to (6.7) is

$$\ell_F(\theta,\pi) = \sum_{j=1}^{J}\left\{\sum_{i=1}^{n_j}\log\left[\sum_{H\in S(G_i)}P(Y_i|X_i,H;\theta)P(H;\pi)\right] + \sum_{i=1}^{n_j}\log P(X_i|G_i) + (N_j-n_j)\log\left[\sum_{X,G}\sum_{H\in S(G)}P(X|G)\eta_j(X,G;\theta,\pi)\right]\right\}$$

which can be rewritten as

$$\ell_F(\theta,\pi) = \sum_{i=1}^{n}\log\left[\sum_{H\in S(G_i)}P(Y_i|X_i,H;\theta)P(H;\pi)\right] + \sum_{g}\sum_{k=1}^{n_g}n_{gk}\log\zeta_{gk} + \sum_{j=1}^{n_j}(N_j-n_j)\log\left[\sum_{g}\sum_{k=1}^{n_g}\eta_j(x_{gk},g;\theta,\pi)\zeta_{gk}\right]. \tag{6.8}$$

We introduce Lagrange multipliers λ_g for the constraint that $\sum_{k=1}^{n_g} \zeta_{gk} = 1$, and take the derivative with respect to ζ_{gk}. Then $\{\zeta_{gk}\}$ which maximize the log-likelihood satisfy

$$\frac{n_{gk}}{\zeta_{gk}} + \sum_{j=1}^{J} \frac{(N_j - n_j)\eta_j(x_{gk}, g; \theta, \pi)}{\sum_g \sum_{k=1}^{n_g} \eta_j(x_{gk}, g; \theta, \pi)\zeta_{gk}} + \lambda_g = 0.$$

If we multiply through by ζ_{gk} and sum over k, then we can solve for λ_g:

$$\lambda_g = -\left(n_g + \sum_{j=1}^{J} \frac{(N_j - n_j)\mu_{gj}}{\sum_g \mu_{gj}}\right)$$

where $\mu_{gj} = \sum_{k=1}^{n_g} \eta_j(x_{gk}, g; \theta, \pi)\zeta_{gk}$. Then

$$\zeta_{gk} = \frac{n_{gk}}{n_g + \sum_{j=1}^{J} \frac{(N_j - n_j)(\mu_{gj} - \eta_j(x_{gk}, g; \theta, \pi))}{\sum_g \mu_{gj}}}.$$

Plugging this back into (6.12) gives us the profile log-likelihood

$$\begin{aligned}
\ell_{FP}(\theta, \pi, \{\mu_{gj}\}) &= \sum_{i=1}^{n} \log \left[\sum_{H \in S(G_i)} P(Y_i|X_i, H; \theta)P(H; \pi)\right] + \sum_g \sum_{k=1}^{n_g} n_{gk} \log n_{gk} \\
&- \sum_g \sum_{k=1}^{n_g} n_{gk} \log \left[n_g + \sum_{j=1}^{J} \frac{(N_j - n_j)(\mu_{gj} - \eta_j(x_g k, g; \theta, \pi))}{\sum_g \mu_{gj}}\right] \\
&+ \sum_{j=1}^{n_j} (N_j - n_j) \log \sum_g \mu_{gj}
\end{aligned}$$

or, up to a constant,

$$\begin{aligned}
\ell_{FP}(\theta, \pi, \{\mu_{gj}\}) &= \sum_{i=1}^{n} \log \left[\sum_{H \in S(G_i)} P(Y_i|X_i, H; \theta)P(H; \pi)\right] \\
&- \sum_{i=1}^{n} \sum_g I(G_i = g) \log \left[n_g + \sum_{j=1}^{J} \frac{(N_j - n_j)(\mu_{gj} - \eta_j(X_i, G_i; \theta, \pi))}{\sum_g \mu_{gj}}\right] \\
&+ \sum_{j=1}^{n_j} (N_j - n_j) \log \sum_g \mu_{gj}. \qquad (6.9)
\end{aligned}$$

Using a similar derivation, we can profile (6.8) over $P(X|G)$ to get (up to a con-

stant)

$$\ell_{CP}(\theta,\pi,\{\mu_{gj}\}) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{J}\delta_{ij}p_j + \log\sum_{H\in S(G_i)} P(Y_i|X_i,H;\theta)P(H;\pi)\right]$$
$$-\sum_{i=1}^{n}\sum_{g} I(G_i=g)\log\left[\frac{n\sum_{j=1}^{J}p_j\eta_j(X_i,G_i;\theta,\pi)}{\sum_{j=1}^{J}\sum_{g}p_j\mu_{gj}} - n_g + \frac{n\sum_{j=1}^{J}p_j\mu_{gj}}{\sum_{j=1}^{J}\sum_{g}p_j\mu_{gj}}\right]$$
$$-n\log\sum_{j=1}^{J}p_j\mu_{gj}. \tag{6.10}$$

These likelihoods can be maximized using Newton-Raphson algorithms, and the covariance can be estimated with sandwich estimators.

If we assume that X and G are independent, then the log-likelihoods for the full and conditional likelihoods respectively are:

$$\ell_F(\theta,\pi) = \sum_{i=1}^{n}\log\left[\sum_{H\in S(G_i)} P(Y_i|X_i,H;\theta)P(H;\pi)\right] + \sum_{i=1}^{n}\log P(X_i)$$
$$+\sum_{j=1}^{J}(N_j-n_j)\log\left[\sum_{X,G}P(X)\eta_j(X,G;\theta,\pi)\right]$$

and

$$\ell_C(\theta,\pi) = \sum_{i=1}^{n}\left[\log\sum_{j=1}^{J}\delta_{ij}p_j + \sum_{H\in S(G_i)} P(Y_i|X_i,H;\theta)P(H;\pi) + \log P(X_i)\right]$$
$$-n\log\sum_{j=1}^{J}\sum_{X,G}\sum_{H\in S(G)} p_j\eta_j(X,G;\theta,\pi)P(X).$$

Suppose X takes K distinct observed values x_k, with probability ζ_k. Let n_{+k} denote the number of times x_k is observed in the dataset. Then we can rewrite ℓ_F as

$$\ell_F(\theta,\pi) = \sum_{i=1}^{n}\log\left[\sum_{H\in S(G_i)} P(Y_i|X_i,H;\theta)P(H;\pi)\right] + \sum_{k=1}^{K} n_{+k}\log\zeta_k$$
$$+\sum_{j=1}^{J}(N_j-n_j)\log\left[\sum_{k=1}^{K}\sum_{g}\eta_j(x_k,g;\theta,\pi)\zeta_k\right].$$

By taking the derivative with respect to ζ_k and introducing a Lagrange multiplier λ for the constraint that $\sum_k \zeta_k = 1$, we obtain

$$\frac{n_{+k}}{\zeta_k} + \sum_{j=1}^{J}(N_j - n_j)\frac{\sum_g \eta_j(x_k, g; \theta, \pi)}{\sum_{k=1}^{K}\sum_g \eta_j(x_k, g; \theta, \pi)\zeta_k} + \lambda = 0.$$

If we then multiply by ζ_k and sum over k we find that $\lambda = -N$. Define $\mu_j = \sum_k \sum_g \eta_j(x_k, g; \theta, \pi)\zeta_k$. Then

$$\zeta_k = \frac{n_{+k}}{\sum_{j=1}^{J}\left[N_j - (N_j - n_j)\frac{\sum_g \eta_j(x_k, g; \theta, \pi)}{\mu_j}\right]}$$

and plugging this back into the log-likelihood we see that the objective function, up to a constant, to be maximized is:

$$\ell_{FP}(\theta, \pi, \{\mu_j\}) = \sum_{i=1}^{n}\left[\log \sum_{H \in S(G_i)} P(Y_i|X_i, H; \theta)P(H; \pi)\right.$$
$$\left. - \log \sum_{j=1}^{J}\left\{N_j - (N_j - n_j)\frac{\sum_H P(Y_i \in \mathcal{C}_j|X_i, H; \theta)P(H; \pi)}{\mu_j}\right\}\right] + \sum_{j=1}^{J}(N_j - n_j)\log \mu_j. \tag{6.11}$$

In this situation, a similar derivation to that presented in Breslow et al. (2003) will show that the observed profile information is a consistent and efficient estimator of the covariance. They require p_j, the strata selection probabilities, to be nonzero for all strata for their Proposition 2.2 which derives the efficient score. The requirement is primarily to ensure invertibility of matrices. This derivation is based on results of Robins et al. (1995); however, they give an alternate explicit computation of the scores which does not have this requirement. Otherwise we simply substitute $P(Y, G|X; \theta) = \sum_{H \in S(G)} P(Y|H, X; \theta)P(H; \pi)$ for $f(y|x; \theta)$ in their derivation in order to get the desired result. Calculations for a Newton-Raphson algorithm to maximize the profile likelihood, and to compute the observed profile information matrix are given in the next section.

For the conditional likelihood, using the same notation of x_k and ζ_k, we find that

taking the derivative under the constraint of $\sum_k \zeta_k = 1$ gives us the following equation

$$\frac{n_{+k}}{\zeta_k} - \frac{n \sum_{j=1}^{J} \sum_g p_j \eta_j(x_k, g; \theta, \pi)}{\sum_{j=1}^{J} p_j \mu_j} + \lambda = 0,$$

where μ_j is defined in the same way as for the previous case. Solving as before leads to

$$\zeta_k = \frac{n_{+k} \sum_{j=1}^{J} p_j \mu_j}{n \sum_{j=1}^{J} p_j \sum_g \eta_j(x_k, g; \theta, \pi)}$$

and a profile likelihood of

$$\ell_{CP}(\theta, \pi) = \sum_{i=1}^{n} \left[\log \sum_{j=1}^{J} \delta_{ij} p_j + \log \sum_{H \in S(G_i)} P(Y_i | X_i, H; \theta) P(H; \pi) \right.$$
$$\left. - \log \sum_{j=1}^{J} p_j \sum_g \eta_j(X_i, g; \theta, \pi) \right]. \tag{6.12}$$

which, up to a constant, is equivalent to the likelihood:

$$\prod_{i=1}^{n} \frac{\sum_{j=1}^{J} p_j \delta_{ij} \sum_{H \in S(G_i)} P(Y_i | X_i, H; \theta) P(H; \pi)}{\sum_{j=1}^{J} p_j \sum_H P(Y_i \in \mathcal{C}_j | X_i, H; \theta) P(H; \pi)} = \prod_{i=1}^{n} P(Y_i, G_i | X_i, R_i = 1). \tag{6.13}$$

Since this takes the form of another conditional likelihood, the covariance matrix of the maximum likelihood estimators can be estimated by the inverse information matrix for this profile likelihood. In order to maximize the likelihood in (6.14) we use a Newton-Raphson algorithm similar to that presented for the maximization of (6.4). Note that the only difference in these two likelihoods is the presence of X_i in the conditional density function $P(Y_i | X_i, H; \theta)$. Define $\tilde{\beta} = (\beta, \gamma)$, $\tilde{Z}_i(h_k, h_l) = (Z(h_k, h_l), X_i)$, and $\tilde{\vartheta} = (\tilde{\beta}, v, \eta)$. Then we can directly apply the algorithm outlined in Section 6.3 to maximize this likelihood by replacing $\beta, Z(h_k, h_l)$, and ϑ by these new versions.

6.5 Newton-Raphson Algorithm to Maximize (6.12)

Under the linear regression model with strata defined as (L_j, U_j), the profile log-likelihood corresponding to (6.12) is

$$\sum_{i=1}^{n}\left[\log\sum_{(h_k,h_l)\in S(G_i)}(2\pi\sigma^2)^{-1/2}\exp\left\{-(Y_i-\beta^T Z(h_k,h_l)-\gamma X_i)^2/(2\sigma^2)\right\}\pi_k\pi_l\right.$$
$$\left.-\log\sum_{j=1}^{J}\left[N_j-\frac{(N_j-n_j)\sum_{k,l}\left[\Phi\left(\frac{U_j-\beta^T Z(h_k,h_l)-\gamma X_i}{\sigma}\right)-\Phi\left(\frac{L_j-\beta^T Z(h_k,h_l)-\gamma X_i}{\sigma}\right)\right]\pi_k\pi_l}{\mu_j}\right]\right]$$
$$+\sum_{j=1}^{J}(N_j-n_j)\log\mu_j.$$

For notational convenience, denote σ^2 as v and write $\vartheta=(\beta,\gamma,v,\pi)$. Combine β and γ into $\Gamma=(\beta,\gamma)$ and correspondingly write $B_i(h_k,h_l)=(Z(h_k,h_l),X_i)$. Let

$$\begin{aligned}
&Q_{ikl}(\vartheta)=\exp\left\{-(Y_i-\Gamma^T B_i(h_k,h_l))^2/(2v)\right\},\\
&R^L_{iklj}(\vartheta)=\{L_j-\Gamma^T B_i(h_k,h_l)\}/\sqrt{v},\\
&R^U_{iklj}(\vartheta)=\{U_j-\Gamma^T B_i(h_k,h_l)\}/\sqrt{v},\\
&S_i(\vartheta)=\sum_{j=1}^{J}\left[N_j-(N_j-n_j)\frac{\sum_{k,l}\left\{\Phi\left(R^U_{iklj}(\vartheta)\right)-\Phi\left(R^L_{iklj}(\vartheta)\right)\right\}\pi_k\pi_l}{\mu_j}\right].
\end{aligned}$$

Also, let $a^{\otimes 2}=aa^T$ and let ϕ be the standard normal density function. Then

$$\frac{\partial\ell_{FP}(\vartheta)}{\partial v}=-\frac{n}{2v}+\sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)\pi_k\pi_l\frac{(Y_i-\Gamma^T B_i(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\theta)}$$
$$-\sum_{i=1}^{n}\frac{\sum_{j=1}^{J}\left[N_j-(N_j-n_j)\sum_{k,l}\left\{\phi\left(R^U_{iklj}(\vartheta)\right)R^U_{iklj}(\vartheta)-\phi\left(R^L_{iklj}(\vartheta)\right)R^L_{iklj}(\vartheta)\right\}\frac{\pi_k\pi_l}{2v\mu_j}\right]}{S_i(\vartheta)}$$

$$\frac{\partial\ell_{FP}(\vartheta)}{\partial\Gamma}=\sum_{i=1}^{n}\frac{\sum_{(h_k,h_l)\in S(G_i)}\frac{Q_{ikl}(\vartheta)}{v}(Y_i-\Gamma^T B_i(h_k,h_l))B_i(h_k,h_l)}{\sum_{(h_k,h_l)\in S(G_i)}Q_{ikl}(\vartheta)\pi_k\pi_l}$$
$$-\sum_{i=1}^{n}\frac{\sum_{j=1}^{J}\left[N_j-(N_j-n_j)\sum_{k,l}\left\{\phi\left(R^U_{iklj}(\vartheta)\right)-\phi\left(R^L_{iklj}(\vartheta)\right)\right\}\frac{B_i(h_k,h_l)}{\sqrt{v}\mu_j}\pi_k\pi_l\right]}{S_i(\vartheta)}$$

$$\frac{\partial \ell_{FP}(\vartheta)}{\partial \pi} = \sum_{i=1}^{n} \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) D_\pi^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l}$$
$$+ \sum_{i=1}^{n} \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \Phi\left(R_{iklj}^U(\vartheta)\right) - \Phi\left(R_{iklj}^L(\vartheta)\right) \right\} D_\pi^{kl} \right]}{S_i(\vartheta)}$$

and for $m = 1, \ldots, K$,

$$\frac{\partial \ell_{FP}(\vartheta)}{\partial \mu_m} = -\sum_{i=1}^{n} \frac{(N_m - n_m) \sum_{k,l} \left\{ \Phi\left(R_{iklm}^U(\vartheta)\right) - \Phi\left(R_{iklm}^L(\vartheta)\right) \right\} \pi_k\pi_l/\mu_m^2}{S_i(\vartheta)}$$
$$+ \frac{N_m - n_m}{\mu_m}.$$

As in Section 6.3, let Δ_{ij} be the Kronecker delta and define D_π^{kl} and $M_{\pi\pi}^{kl}$, the first and second derivatives of $\pi_k\pi_l$ as before. Then we can compute the observed profile information matrix by taking the negative of the following second derivatives:

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial v^2} = \frac{n}{2v^2} + \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \left\{ \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^4}{4v^4} - \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^2}{v^3} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\}^2 \right]$$
$$- \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi(R_{iklj}^U(\vartheta)) \left(\left(R_{iklj}^U(\vartheta)^3\right) - 3R_{iklj}^U(\vartheta) \right) \right\} \frac{\pi_k\pi_l}{4v^2\mu_j} \right]}{S_i(\vartheta)} \right.$$
$$- \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi(R_{iklj}^L(\vartheta)) \left(\left(R_{iklj}^L(\vartheta)^3\right) - 3R_{iklj}^L(\vartheta) \right) \right\} \frac{\pi_k\pi_l}{4v^2\mu_j} \right]}{S_i(\vartheta)}$$
$$\left. - \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left(\phi(R_{iklj}^U(\vartheta)) R_{iklj}^U(\vartheta) - \phi(R_{iklj}^L(\vartheta)) R_{iklj}^L(\vartheta) \right) \frac{\pi_k\pi_l}{2v\mu_j} \right]}{S_i(\vartheta)} \right\}^2 \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial v \partial \Gamma} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) B_i(h_k,h_l) \left\{ \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^3}{2v^3} - \frac{(Y_i - \Gamma^T B_i(h_k,h_l))}{v^2} \right\}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right.$$
$$- \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \Gamma^T B_i(h_k,h_l)) B_i(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\}$$
$$\left. \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\} \right]$$
$$- \left[\frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi(R^U_{iklj}(\vartheta)) \left(R^U_{iklj}(\vartheta)^2 - 1 \right) \right\} \frac{\pi_k\pi_l}{2v^{3/2}\mu_j} B_i(h_k,h_l) \right]}{S_i(\vartheta)} \right.$$
$$- \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi(R^L_{iklj}(\vartheta)) \left(R^L_{iklj}(\vartheta)^2 - 1 \right) \right\} \frac{\pi_k\pi_l}{2v^{3/2}\mu_j} B_i(h_k,h_l) \right]}{S_i(\vartheta)}$$
$$- \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left(\phi(R^U_{iklj}(\vartheta)) R^U_{iklj}(\vartheta) - \phi(R^L_{iklj}(\vartheta)) R^L_{iklj}(\vartheta) \right) \frac{\pi_k\pi_l}{2v\mu_j} \right]}{S_i(\vartheta)} \right\}$$
$$\left. \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left(\phi(R^U_{iklj}(\vartheta)) - \phi(R^L_{iklj}(\vartheta)) \right) \pi_k\pi_l \frac{B_i(h_k,h_l)}{\sqrt{v}\mu_j} \right]}{S_i(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial (\Gamma\Gamma^T)} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \left\{ \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^2}{v^2} - v^{-1} \right\} B_i^{\otimes 2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \Gamma^T B_i(h_k,h_l)) B_i(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\}^{\otimes 2} \right]$$
$$- \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \phi(R^U_{iklj}(\vartheta)) R^U_{iklj}(\vartheta) \frac{\pi_k\pi_l}{\mu_j} \left(\frac{B_i(h_k,h_l)}{\sqrt{v}} \right)^{\otimes 2} \right]}{S_i(\vartheta)} \right.$$
$$- \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \phi(R^L_{iklj}(\vartheta)) R^L_{iklj}(\vartheta) \frac{\pi_k\pi_l}{\mu_j} \left(\frac{B_i(h_k,h_l)}{\sqrt{v}} \right)^{\otimes 2} \right]}{S_i(\vartheta)}$$
$$\left. - \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left(\phi(R^U_{iklj}(\vartheta)) - \phi(R^L_{iklj}(\vartheta)) \right) \pi_k\pi_l \frac{B_i(h_k,h_l)}{\sqrt{v}\mu_j} \right]}{S_i(\vartheta)} \right\}^{\otimes 2} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial v \partial \pi^T} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)(D_\pi^{kl})^T \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^2}{2v^2}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right.$$
$$- \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} \frac{(Y_i - \Gamma^T B_i(h_k,h_l))^2}{2v^2} Q_{ikl}(\vartheta)}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\} \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)(D_\pi^{kl})^T}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\} \Bigg]$$
$$- \sum_{i=1}^{n} \left[\left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \phi\left(R_{iklj}^U(\vartheta)\right) R_{iklj}^U(\vartheta) \frac{\pi_k\pi_l}{2v\mu_j} \right]}{S_i(\vartheta)} \right.$$
$$\left. - \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \phi\left(R_{iklj}^L(\vartheta)\right) R_{iklj}^L(\vartheta) \frac{\pi_k\pi_l}{2v\mu_j} \right]}{S_i(\vartheta)} \right\}$$
$$\left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \Phi\left(R_{iklj}^U(\vartheta)\right) - \Phi\left(R_{iklj}^L(\vartheta)\right) \right\} \frac{D_\pi^{kl}}{\mu_j} \right]}{S_i(\vartheta)} \right\}^T$$
$$\left. + \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi\left(R_{klj}^U(\theta)\right) R_{klj}^U(\theta) - \phi\left(R_{klj}^L(\theta)\right) R_{klj}^L(\theta) \right\} \frac{(D_\pi^{kl})^T}{2v\mu_j} \right]}{S_i(\vartheta)} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial \Gamma \partial \pi^T} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \Gamma^T B_i(h_k,h_l)) B_i(h_k,h_l)}{v} (D_\pi^{kl})^T}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right.$$
$$\left. - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) \frac{(Y_i - \Gamma^T B_i(h_k,h_l)) B_i(h_k,h_l)}{v}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\} \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) D_\pi^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\}^T \right]$$
$$- \sum_{i=1}^{n} \left[\left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi\left(R_{iklj}^U(\vartheta)\right) - \phi\left(R_{iklj}^L(\vartheta)\right) \right\} \frac{B_i(h_k,h_l)}{\sqrt{v}\mu_j} \pi_k\pi_l \right]}{S_i(\vartheta)} \right\}$$
$$\left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \Phi\left(R_{iklj}^U(\vartheta)\right) - \Phi\left(R_{iklj}^L(\vartheta)\right) \right\} \frac{D_\pi^{kl}}{\mu_j} \right]}{S_i(\vartheta)} \right\}^T$$
$$\left. + \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi\left(R_{iklj}^U(\vartheta)\right) - \phi\left(R_{iklj}^L(\vartheta)\right) \right\} \frac{B_i(h_k,h_l)}{\sqrt{v}\mu_j} (D_\pi^{kl})^T \right]}{S_i(\vartheta)} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial \pi \partial \pi^T)} = \sum_{i=1}^{n} \left[\frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) M_{\pi\pi}^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} - \left\{ \frac{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta) D_{\pi}^{kl}}{\sum_{(h_k,h_l)\in S(G_i)} Q_{ikl}(\vartheta)\pi_k\pi_l} \right\}^{\otimes 2} \right]$$

$$+ \sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \Phi\left(R_{iklj}^{U}(\vartheta)\right) - \Phi\left(R_{iklj}^{L}(\vartheta)\right) \right\} \frac{M_{\pi\pi}^{kl}}{\mu_j} \right]}{S_i(\vartheta)} \right.$$

$$\left. + \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \Phi\left(R_{iklj}^{U}(\vartheta)\right) - \Phi\left(R_{iklj}^{L}(\vartheta)\right) \right\} \frac{D_{\pi}^{kl}}{\mu_j} \right]}{S_i(\vartheta)} \right\}^{\otimes 2} \right]$$

For $m, n = 1, \ldots, K$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial v \partial \mu_m} = \sum_{i=1}^{n} \left[\frac{(N_m - n_m) \sum_{k,l} \left\{ \phi\left(R_{iklm}^{U}(\vartheta)\right) R_{iklm}^{U}(\vartheta) \right\} \frac{\pi_k\pi_l}{2v\mu_m^2}}{S_i(\vartheta)} \right.$$

$$- \frac{(N_m - n_m) \sum_{k,l} \left\{ \phi\left(R_{iklm}^{L}(\vartheta)\right) R_{iklm}^{L}(\vartheta) \right\} \frac{\pi_k\pi_l}{2v\mu_m^2}}{S_i(\vartheta)}$$

$$+ \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi\left(R_{iklj}^{U}(\vartheta)\right) R_{iklj}^{U}(\vartheta) - \phi\left(R_{iklj}^{L}(\vartheta)\right) R_{iklj}^{L}(\vartheta) \right\} \frac{\pi_k\pi_l}{2v\mu_j} \right]}{S_i(\vartheta)} \right\}$$

$$\left. \left\{ \frac{(N_m - n_m) \sum_{k,l} \left\{ \Phi\left(R_{iklm}^{U}(\vartheta)\right) - \Phi\left(R_{iklm}^{L}(\vartheta)\right) \right\} \frac{\pi_k\pi_l}{\mu_m^2}}{S_i(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial \Gamma \partial \mu_m} = \sum_{i=1}^{n} \left[\frac{(N_m - n_m) \sum_{k,l} \left\{ \phi\left(R_{iklm}^{U}(\vartheta)\right) - \phi\left(R_{iklm}^{L}(\vartheta)\right) \right\} \frac{\pi_k\pi_l B_i(h_k,h_l)}{\sqrt{v}\mu_m^2}}{S_i(\vartheta)} \right.$$

$$+ \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \phi\left(R_{iklj}^{U}(\vartheta)\right) - \phi\left(R_{iklj}^{L}(\vartheta)\right) \right\} \frac{\pi_k\pi_l B_i(h_k,h_l)}{\sqrt{v}\mu_j} \right]}{S_i(\vartheta)} \right\}$$

$$\left. \left\{ \frac{(N_m - n_m) \sum_{k,l} \left\{ \Phi\left(R_{iklm}^{U}(\vartheta)\right) - \Phi\left(R_{iklm}^{L}(\vartheta)\right) \right\} \frac{\pi_k\pi_l}{\mu_m^2}}{S_i(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial \pi \partial \mu_m} = -\sum_{i=1}^{n} \left[\frac{(N_m - n_m) \sum_{k,l} \left\{ \Phi\left(R_{iklm}^{U}(\vartheta)\right) - \Phi\left(R_{iklm}^{L}(\vartheta)\right) \right\} \frac{D_{\pi}^{kl}}{\mu_m^2}}{S_i(\vartheta)} \right.$$

$$+ \left\{ \frac{\sum_{j=1}^{J} \left[N_j - (N_j - n_j) \sum_{k,l} \left\{ \Phi\left(R_{iklj}^{U}(\vartheta)\right) - \Phi\left(R_{iklj}^{L}(\vartheta)\right) \right\} \frac{D_{\pi}^{kl}}{\mu_j} \right]}{S_i(\vartheta)} \right\}$$

$$\left. \left\{ \frac{(N_m - n_m) \sum_{k,l} \left\{ \Phi\left(R_{iklm}^{U}(\vartheta)\right) - \Phi\left(R_{iklm}^{L}(\vartheta)\right) \right\} \frac{\pi_k\pi_l}{\mu_m^2}}{S_i(\vartheta)} \right\} \right]$$

$$\frac{\partial^2 \ell_{FP}(\vartheta)}{\partial \mu_m \partial \mu_n} = \sum_{i=1}^{n} \left[\frac{2(N_m - n_m)\Delta_{mn} \sum_{k,l} \left\{ \Phi\left(R^U_{iklm}(\vartheta)\right) - \Phi\left(R^L_{iklm}(\vartheta)\right) \right\} \frac{\pi_k \pi_l}{\mu_m^3}}{S_i(\vartheta)} \right.$$
$$- \left\{ \frac{(N_m - n_m) \sum_{k,l} \left\{ \Phi\left(R^U_{iklm}(\vartheta)\right) - \Phi\left(R^L_{iklm}(\vartheta)\right) \right\} \frac{\pi_k \pi_l}{\mu_m^2}}{S_i(\vartheta)} \right]$$
$$\left. \left\{ \frac{(N_n - n_n) \sum_{k,l} \left\{ \Phi\left(R^U_{ikln}(\vartheta)\right) - \Phi\left(R^L_{ikln}(\vartheta)\right) \right\} \frac{\pi_k \pi_l}{\mu_n^2}}{S_i(\vartheta)} \right\} \right]$$
$$- \frac{\Delta_{mn}(N_m - n_m)}{\mu_m^2}$$

6.6 Simulation Studies

We investigated the performance of the proposed methods under a variety of selective genotyping designs through simulation studies. We considered all the designs studied in this chapter; namely, designs both with and without covariates, and both with and without stratum information retained. In accordance with the two-phase design described earlier, we generated individuals from the joint distribution of the trait value, genotype, and covariates, and selected a subset of 500 for genotyping based on selection probabilities for three strata. The covariates were discarded for a subset of simulations, and by either retaining or discarding the stratum information for all individuals, we could apply the full and conditional likelihood methods. Similarly to the previous chapter, this provided information about the relative efficiency of the two methods. These methods were compared to standard methods based on the prospective likelihood which is equivalent to (6.7) when $N_j = n_j$, $\forall j$.

The first study was concerned with the performance of the two-phase study design when no covariates were measured. Trait values were generated for a single SNP model under model (6.1) with $\alpha = 0$, $\sigma^2 = 1$, and $\beta = 0, 0.1, 0.2$ and 0.3. The potential causal variant was the minor allele, which was generated with frequency 0.1 for the additive mode of inheritance and 0.3 for the recessive mode of inheritance. In the previous chapter we saw very little difference between the additive and dominant models,

hence our focus on the additive and recessive modes of inheritance. To evaluate the differences in performance due to the more general design, we considered effects of varying selection cutpoints and probabilities. The effects of varying MAF and effect size were already well characterized and carry over to the two-phase design.

As we would expect from the results in Chapter 5, both the full and conditional likelihoods provide unbiased estimates of the haplotype effect and correct type I error. The standard error estimates agree with the true variation, and the confidence intervals have correct coverage. The conditional likelihood has negligible loss of power and nearly identical results to the full likelihood, so we present only the results for the full likelihood here. The prospective likelihood preserves the Type I error, and has similar power to the proposed methods when selection occurs symmetrically. However, many of the aspects noted earlier, such as high bias, low coverage, and reduced power for asymmetric sampling, apply under the new design. Figure 6.1 shows the drastic drop in coverage probability for the prospective method as the haplotype effect size increases. This is due to increasing bias in haplotype effect estimation.

We consider a design with three strata, which is comparable to the previous results when the selection probability is zero for the middle stratum. The two cutpoints for this design correspond to c_L and c_U from the previous chapter. In Figure 6.2 we demonstrate the reduction in power for the prospective method as we increase asymmetricity in the cutpoints, and fix the selection probabilities to be equal in the two tails, and zero otherwise. The additive model has much higher power than the recessive model, but there is relatively little difference between the full and prospective likelihoods for this mode of inheritance. In contrast, for the recessive mode of inheritance, there can be as much as a 25% drop in power.

For this design, we also consider the effects of varying the selection probabilities, while keeping the cutpoints fixed. We would expect this to have a similar effect to changing the cutoffs, since in both situations we genotype a larger sample from one tail than the other. The choice in practice of which parameter to modify will depend on prior information about the population and ease of recruitment for a given strategy.

Figure 6.3 shows the additive and recessive models, and again the most obvious feature is the reduced power for the prospective analysis relative to the full analysis under the recessive model.

With multiple strata, however, we may allow for positive selection in all strata rather than only in the tails. This results in decreased power for all methods, since the selection is not as extreme; it may be necessary in practice if it is impossible to sample sufficient individuals from some strata. Figure 6.4 illustrates this reduction in power for the full likelihood method. We compare a three strata model with selection probabilities of 0.5, 0.05, and 0.5, when the cutpoints are balanced at -1 and 1, to a model with zero selection probability for the middle stratum. The drop in power is relatively low and stays below 6% as the haplotype effect size increases. As the proportion of individuals selected from the stratum increases, the design approaches a simple random sample, and we see this effect in Figure 6.1 with the coverage probability. The prospective likelihood has reduced bias and improved coverage when there is positive sampling from all strata, although the coverage is still far below nominal levels.

The second study considers the scenario where a single binary covariate is measured. Genotypes and trait were generated as in the first study. We assume independence between covariates and genotypes, and generate X according to a Bernoulli distribution with probability 0.1. The conclusions are essentially the same as described previously, with the addition that the results for the prospective method regarding haplotype effects generalize to the covariate effect as well. The full and conditional methods estimate the covariate effect without bias and with similar power.

6.7 Discussion

We present several new methods for use in selective genotyping designs in this chapter. The designs considered are more general than have been previously studied, and the proposed full and conditional likelihoods are shown to work well under a range

of conditions. In theory, the full likelihood for design 1 from Chapter 5 is applicable to the two-phase setup. This incorporates additional information by retaining actual trait values rather than stratum information for individuals who are not genotyped. However, in both this and the previous chapter we saw that there is very little loss of efficiency between the full and conditional approaches, so in practice one may simply use the conditional likelihood and discard both the additional trait and stratum values.

Lawless et al. (1999) suggest that there are situations in which the stratum sizes may contain substantial additional information; this was not observed in our simulations. Maximum likelihood performs well in all their simulations, but increasing the number of strata appears to slow down achievement of asymptotic optimality. These are issues which we intend to explore further. Our simulations obviously do not encompass all practical situations, but illustrate properties of the methods when varying some basic features of association studies.

The designs considered here may be extended further. We assume that the only information observed if a subject is not genotyped is the stratum containing that subject's trait value. In practice, however, covariates may be observed for an individual even if they are not genotyped, since the cost of measuring environmental factors is generally much less than for genotyping. Lawless et al. (1999) state that the full likelihood cannot be maximized exactly when both the trait and covariates are continuous. However, Robins et al. (1995) propose an adaptive semiparametric efficient estimator when the vector of covariates has at most two continuous components. Extending their methods would allow us to consider designs where the trait values and covariates are measured on the full set of N individuals from the population, as well as on the subset of n genotyped individuals.

Finally, in the near future we intend to present theoretical justification for using the observed profile information matrix to compute standard errors when covariates and genotype are not assumed to be independent. For this we need to show that the design satisfies the conditions of Murphy and van der Vaart (2000), as is done for the general two-phase design in Breslow et al. (2003).

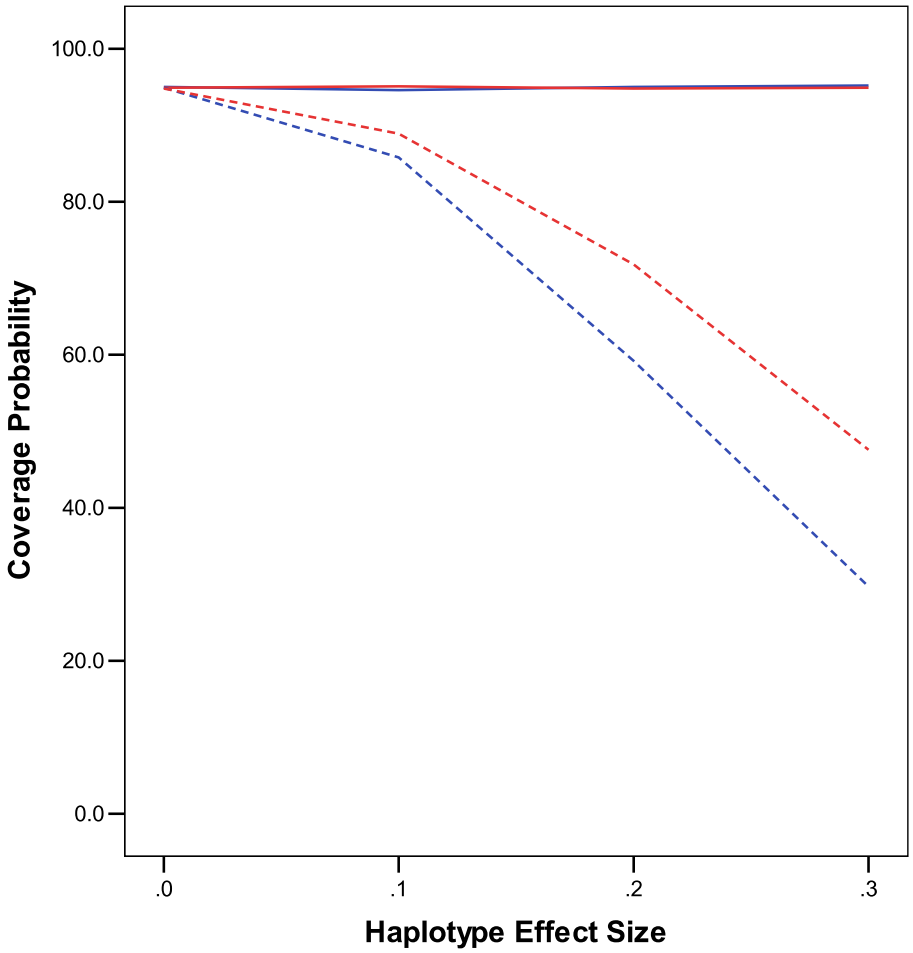


Figure 6.1: Empirical coverage probability for the 95% confidence interval for a 3-strata, 1-SNP model as a function of the haplotype effect size. All curves are generated under an additive model with MAF=0.1. The red curves correspond to selection probabilities of 0.5, 0.05, and 0.5, and the blue curves correspond to selection probabilities of 0.5, 0.05, and 0.5. Solid curves pertain to the full likelihood analysis, while dotted curves pertain to the prospective analysis. The cutpoints for the strata are -1 and 1. All values are based on 10,000 iterations.

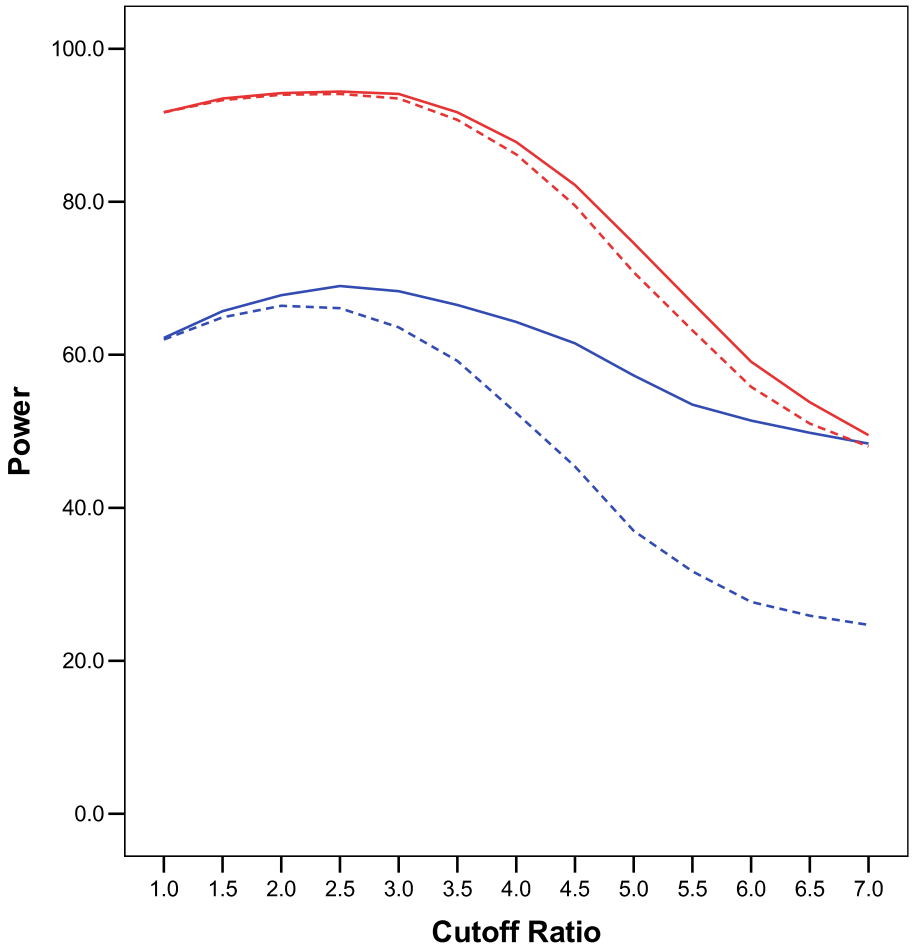


Figure 6.2: Empirical power for a 3-strata, 1-SNP model as a function of the absolute ratio of the lowest cutpoint to the highest cutpoint. The highest cutpoint is set at 0.4. The red curves correspond to an additive model with MAF=0.1, and the blue curves correspond to a recessive model with MAF=0.3. All models had effect size $\beta = .3$. Solid curves pertain to the full likelihood analysis, while dotted curves pertain to the prospective analysis. The selection probabilities for the three strata are .5, 0 and .5 respectively. All values are based on 10,000 iterations.

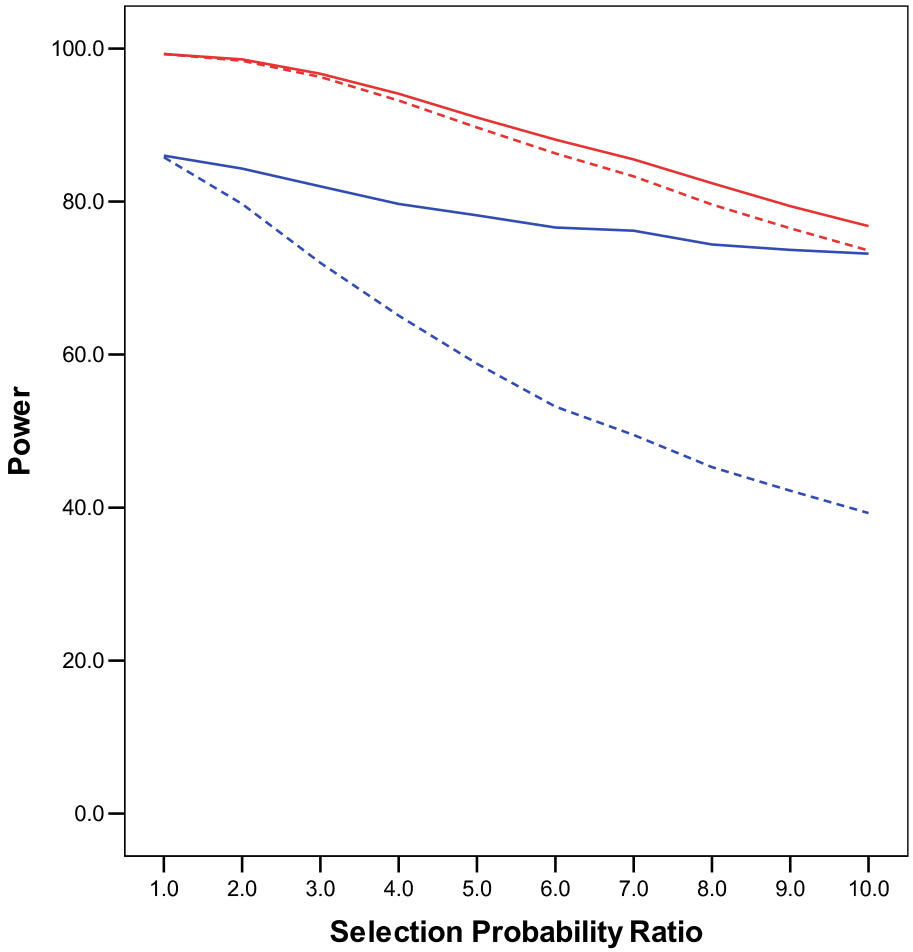


Figure 6.3: Empirical power for a 3-strata, 1-SNP model as a function of the ratio of selection probabilities for the highest to the lowest strata. The selection probability for the lowest stratum is set at 0.1, and the middle stratum has zero selection probability. The red curves correspond to an additive model with MAF=0.1, and the blue curves correspond to a recessive model with MAF=0.3. All models had effect size $\beta = .3$. Solid curves pertain to the full likelihood analysis, while dotted curves pertain to the prospective analysis. The cutoffs for the strata are -1 and 1. All values are based on 10,000 iterations.

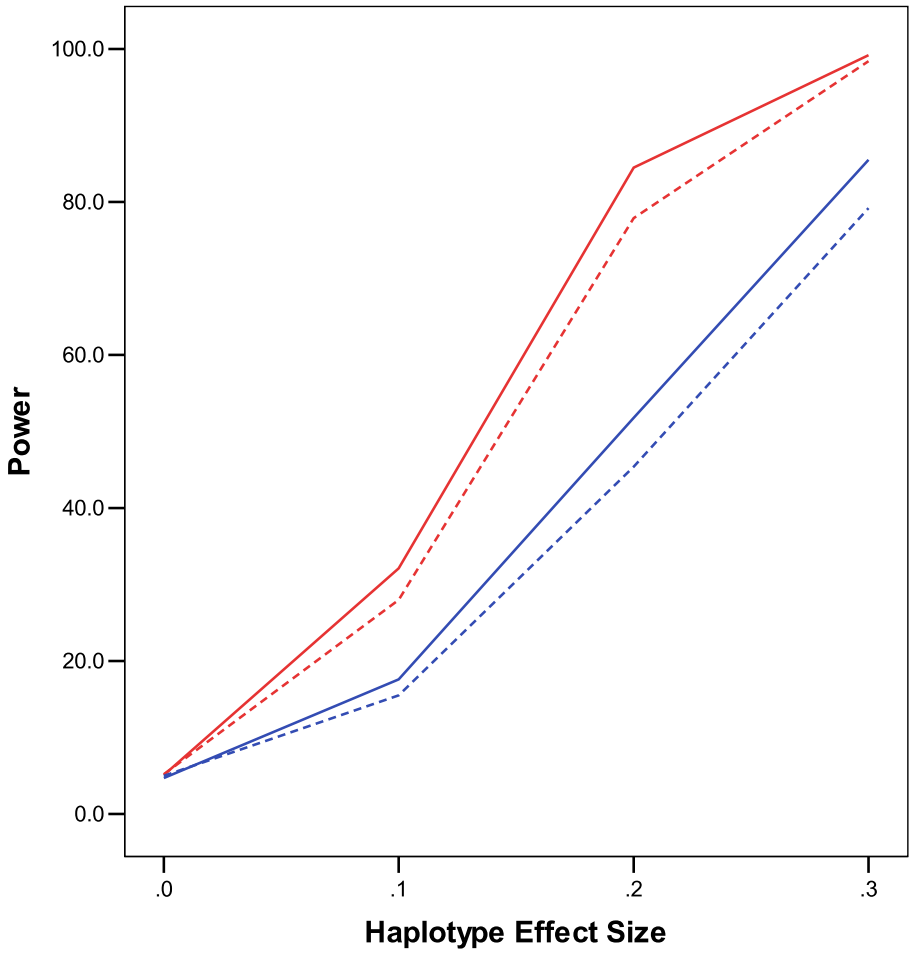


Figure 6.4: Empirical power for a 3-strata, 1-SNP model as a function of the haplotype effect size. The red curves correspond to an additive model with MAF=0.1, and the blue curves correspond to a recessive model with MAF=0.3. All results pertain to the full likelihood analysis. Solid curves are for selection probabilities of 0.5, 0, and 0.5, while dotted curves are for selection probabilities of 0.5, 0.05 and 0.5. The cutpoints for the strata are -1 and 1. All values are based on 10,000 iterations.

REFERENCES

Akey J, Jin L, Xiong M. 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**:291-300.

Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, Healey LA, Kaplan SR, Liang MH, Luthra HS, et al. 1998. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis and Rheumatism* **31**:315-324.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**:781-791.

Breslow N, McNeney B, Wellner JA. 2003. Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Annals of Statistics* **31**:1110-1139.

Browning SR. 2006. Multilocus association mapping using variable-length Markov chains. *American Journal of Human Genetics* **78**:903-913.

Chen HY. 2003. A note on the prospective analysis of outcome-dependent samples. *Journal of the Royal Statistical Society B* **65**:575-584.

Chen L, Storey JD. 2006. Relaxed Significance Criteria for Linkage Analysis. *Genetics* **173**:2371-2381.

Chen Z, Zheng G, Ghosh K, Li Z. 2005. Linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *American Journal of Human Genetics* **77**:661-669.

Collins FS, Guyer MS, Chakravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**:1580-1581.

Cordell HJ. 2006. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genetic Epidemiology* **30**:259-275.

Cornish KM, Manly T, Savage R, Swanson J, Morisano D, Butler N, Grant C, Cross G, Bentley L, Hollis CP. 2005. Association of the dopamine transporter (DAT1) 10/10-repeat genotype with ADHD symptoms and response inhibition in a general populations sample. *Molecular Psychiatry* **10**:686-698.

De Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nature Genetics* **37**:1217-1223.

Dempster AP, Rubin DB. 1983. Introduction. In *Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography* (eds. WG Madow, I Olkin, DB Rubin), pp. 3-10. New York: Academic Press

Dudbridge F, Koeleman BPC. 2004. Efficient computation of significance levels for multiple association in large studies of correlated data, including genomewide association studies. *American Journal of Human Genetics* **75**:424-435.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *American Journal of Human Genetics* **75**:35-43.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73**:1316-1329.

Excoffier L, Laval G, Balding D. 2003. Gametic phase estimation over large genomic regions using an adaptive window approach. *Human Genomics* **1**:7-19.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**:921-927.

Fallin D, Schork NJ. 2000. Accuracy of haplotype frequency estimation for bialleli loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics* **67**:947-959.

French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM. 2006. Simple estimates of haplotype relative risks in case-control data. *Genetic Epidemiology* **30**:485-494.

Garner C. 2007. Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology* **31**:288-295.

Ge Y, Dudoit S, Speed TP. 2003. Resampling-based multiple testing for microarray data analysis (with discussion). *Test* **12**:1-77.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, Huang W, Liu B, Shen Y, et al. 2003. The International HapMap Project. *Nature* **426**:789-796.

Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, et al. 2006. A common genetic variant is associated with adult and childhood obesity. *Science* **312**:279-283.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**:95-108.

Huang BE, Lin DY. 2007. Efficient association mapping of quantitative trait loci with selective genotyping. *American Journal of Human Genetics* **80**:567-576.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**:1299-1320.

Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**:65-70.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**:337-338.

Jawaheer D, Lum RF, Amos CI, Gregersen PK, Criswell LA. 2004. Clustering of disease features within 512 multicase rheumatoid arthritis families. *Arthritis and Rheumatism* **50**:736-741.

Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Monteiro J, Kern M, Criswell LA, Albani S, Nelson JL, et al. 2001. A genome-wide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *American Journal of Human Genetics* **68**:927-936.

Jimenez-Sanchez G, Childs B, Valle D. 2001. Human disease genes. *Nature* **409**:853-855.

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et. al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**:385-389.

Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I. 2005. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genetic Epidemiology* **28**:261-272.

Laitinen T, Kauppi P, Ignatius J, Ruotsalainen T, Daly MJ, Kääriäinen H, Kruglyak L, Laitinen H, de la Chapelle A, Lander ES, et al. 1997. Genetic control of serum IgE levels and asthma: linkage and linkage disequilibrium studies in an isolated population. *Human Molecular Genetics* **6**:2069-2076

Lawless JF, Kalbfleisch JD, Wild CJ. 1999. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society B* **61**:413-438.

Lehmann EL, Romano JP. 2005. Generalizations of the familywise error rate. *Annals of Statistics* **33**:1138-1154.

Li Y, Sung W-K, Liu JJ. 2007. Association mapping via regularized regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. *American Journal of Human Genetics* **80**:705-715.

Lin DY. 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**:781-787.

Lin DY, Huang BE. 2007. The use of inferred haplotypes in downstream analyses. *American Journal of Human Genetics* **80**:577-579.

Lin DY, Zeng D. 2006. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**:89-118.

Lin DY, Zeng D, Millikan R. 2005. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology* **29**:299-312.

Lin S, Chakravarti A, Cutler DJ. 2004. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics* **36**:1181-1188.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PVK, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. *American Journal of Human Genetics* **77**:685-693.

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, et al. 2006. A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics* **78**:437-450.

Mathias RA, Gao P, Goldstein JL, Wilson AF, Pugh EW, Furbert-Harris P, Dunson GM, Malveaux FJ, Togias A, Barnes KC, et al. 2006. A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. *BMC Genetics* **7**:38.

Mensah EK, Gilthorpe MS, Davies CF, Keen LJ, Adamson PJ, Roman E, Morgan GJ, Bidwell JL, Law GR. Haplotype uncertainty in association studies. *Genetic Epidemiology* **31**:348-357.

Merriman TR, Cordell HJ, Eaves IA, Danoy PA, Coraddu F, Barber R, Cucca F, Broadley S, Sawcer S, Compston A, et al. 2001. Suggestive evidence for association of human chromosome 18q12-q21 and its orthologue on rat and mouse chromosome 18 with several autoimmune diseases. *Diabetes* **50**:184-194.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *American Journal of Human Genetics* **74**:945-953.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology* **23**:221-233.

Murphy SA, Van der Vaart AW. 2000. On profile likelihood. *Journal of the American Statistical Association* **95**:449-485.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics* **70**:157-169.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32**:650-654.

Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW, Wolfe F, Kastner DL, Alfredsson L, Altshuler D, et al. 2005. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with *PTPN22*, *CTLA4* and *PADI4*. *American Journal of Human Genetics* **77**:1044-1060.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* **66**:403-411.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends in Genetics* **17**:502-510.

Risch NJ. 2000. Searching for genetic determinants in the new millenium. *Nature* **405**:847-856.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273**:1616-1617.

Robins JM, Hsieh F, Newey W. 1995. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society B* **57**:409-424.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**:425-434.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**:629-644.

Scott AJ, Wild CJ. 1997. Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**:57-71.

Scott AJ, Wild CJ. 2000. Maximum likelihood for generalised case-control studies. Statistical design of medical experiments. II. *Journal of Statistical Planning and Inference* **96**:3-27.

Slatkin M. 1999. Disequilibrium mapping of a quantitative-trait locus in an expanding population. *American Journal of Human Genetics* **64**:1765-1773

Spinka C, Carroll RJ, Chatterjee N. 2005. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* **29**:108-127.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**:978-989.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **72**:1162-1169.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**:449-462.

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity* **55**:179-190.

Thomas DC, Haile RW, Duggan D. 2005. Recent developments in genomewide association scans: a workshop summary and review. *American Journal of Human Genetics* **77**:337-345.

Tzeng J-Y, Wang C-H, Kao J-T, Hsiao CK. 2006. Regression-Based Association Analysis with Clustered Haplotypes through Use of Genotypes. *American Journal of Human Genetics* **78**:231-242.

van Gestel S, Houwing-Duistermaat JJ, Adolfsson R, van Duijn CM, van Broeckhoven C. 2000. Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior Genetics* **30**:141-146

Wallace C, Chapman JM, Clayton DG. 2006. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *American Journal of Human Genetics* **78**:498-504.

Wang WYS, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* **6**:109-118.

Weir BS. 1996. *Genetic Data Analysis II*. Sinauer Associates, Inc. Sunderland, MA.

Westfall PH, Young SS. 1993. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, New York.

Wiuf C, Posada D. 2003. A coalescent model of recombination hotspots. *Genetics* **164**:407-417.

Xiong M, Fan R, Jin L. 2002. Linkage disequilibrium mapping of quantitative trait loci under truncation selection. *Human Heredity* **53**:158-172

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* **53**:79-91.

Zhao LP, Li SS, Khalid N. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics* **72**:1231-1250.

Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. 2002. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* **58**:413-421.